Region-Based Image Clustering and Retrieval Using Multiple Instance Learning

Chengcui Zhang and Xin Chen

Department of Computer and Information Sciences,
University of Alabama at Birmingham
{zhang, chenxin}@cis.uab.edu

Abstract. Multiple Instance Learning (MIL) is a special kind of supervised learning problem that has been studied actively in recent years. We propose an approach based on One-Class Support Vector Machine (SVM) to solve MIL problem in the region-based Content Based Image Retrieval (CBIR). This is an area where a huge number of image regions are involved. For the sake of efficiency, we adopt a Genetic Algorithm based clustering method to reduce the search space. Relevance Feedback technique is incorporated to provide progressive guidance to the learning process. Performance is evaluated and the effectiveness of our retrieval algorithm is demonstrated in comparative studies.

1 Introduction

Relevance feedback (RF) technique is widely used to incorporate user's concept with the learning process [2][4] for Content-Based Image Retrieval (CBIR). Most of the existing RF-based approaches consider each image as a whole, which is represented by a vector of N dimensional image features. However, user's query interest is often just one part of the query image. Therefore it is more reasonable to view it as a set of semantic regions. In this context, the goal of image retrieval is to find the semantic region(s) of user's interest. Since each image is composed of several regions and each region can be taken as an instance, region-based CBIR is then transformed into a Multiple Instance Learning (MIL) problem [5]. Maron et al. applied MIL into natural scene image classification [5]. Each image is viewed as a bag of semantic regions (instances). In the scenario of MIL, the labels of individual instances in the training data are not available, instead the bags are labeled. When applied to RF-based CBIR, this corresponds to the scenario that the user gives feedback on the whole image (bag) although he/she may be only interested in a specific region (instance) of that image.

In order to support region-based image retrieval, we need to divide each image into several semantic regions (instances). However, this further increases the search. Given the huge amount of semantic regions in this problem, we first preprocess image regions by dividing them into clusters. In this way the search space can be reduced to a few clusters that are relevant to the query region. K-means is a traditional clustering method and has been widely used in image clustering. However, it is incapable of finding non-convex clusters and tends to fall into local optimum especially when the

W.-K. Leow et al. (Eds.): CIVR 2005, LNCS 3568, pp. 194–204, 2005.

number of data objects is large. In contrast, Genetic algorithm [9] is known for its robustness and ability to approximate global optimum. In this study, we adapted it to suit our needs of clustering image regions.

After clustering, our proposed system applies MIL to learn the region of interest from users' relevance feedback on the whole image. In particular, the proposed learning algorithm concentrates on the positive bags (images). The motivation is that positive samples are all alike, while negative samples are each bad in their own way. Instead of building models for both positive class and negative class, it makes more sense to assume that all positive regions are in one class while the negative regions are outliers of the positive class. Therefore, we applied One-Class Support Vector Machine (SVM) [1] to solve the MIL problem in CBIR. Chen et al. [6] and Gondra [10] use One-Class SVM in image retrieval but, again, it is applied to the image as a whole. In our approach, One-Class SVM is used to model the non-linear distribution of image regions. Each region of the test images is given a similarity score by the evaluation function built from the model. The images with the highest scores are returned to the user as query results. However, the critical issue here is how to transform the traditional SVM learning, in which labeled training instances are readily available, to a MIL learning problem where only the labels of bags (e.g. images with positive/negative feedbacks) are available. In this study, we proposed a method to solve the aforementioned problem and our experiments show that high retrieval accuracy can be achieved usually within 4 iterations.

In Section 2, we present the clustering method. In Section 3, the detailed learning and retrieval approach is discussed. In Section 4, the overall system is illustrated and the experimental results are presented. Section 5 concludes the paper.

2 Genetic Algorithm Based Clustering

2.1 Overview of Genetic Algorithm

The basic idea of Genetic Algorithm originates from the theory of evolution -- "survival of the fittest". It was formally introduced in the 1970s by John Holland [8]. The overview of genetic algorithm is shown in Fig. 1.

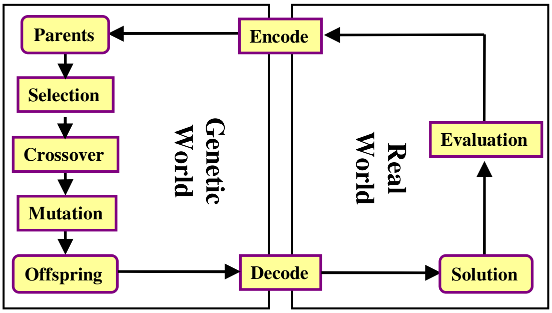


Fig. 1. Genetic Algorithm Overview

The possible solutions to a real world problem are first encoded. Each solution forms a chromosome. A population is a group of chromosomes. From the first generation (parents), these chromosomes will go through *Selection*, *Crossover* and *Mutation* and generate the next generation (offspring). The next generation of chromosomes is decoded back into real world solutions. An objective function is used to measure the fitness of each individual solution. This accomplishes the evolution of the first generation. Genetic algorithm then starts to run the next generation.

2.2 Genetic Algorithm Design for Image Region Clustering

In image region clustering, the target problem is to group semantic image regions into clusters according to their similarities. Each cluster is represented by its centroid. With this objective, we design the genetic algorithm below.

2.2.1 Encoding

A feasible solution to a clustering problem would be a set of centroids. Therefore we give each region an ID: 1, 2, ...,n (n is an integer). The centroids are represented by their ID in the chromosome.

98	56	10	23	65	35	22	469	16

Fig. 2. A Chromosome Example

Fig. 2 is an example of a chromosome. In this chromosome, each integer is a gene in genetic world which corresponds to the ID of a centroid region.

2.2.2 Objective Function

The objective of image region clustering is to find the optimal combination that minimizes the function below:

$$F(R)=\sum_{j=1}^{k}\sum_{i=1}^{n} d(p_i, rep[p_{i,}, R_j]) \tag{1}$$

p_i is an image region in the cluster R_j which is represented by an representative image region $rep[p_i, R_j]$. n is the total number of image regions and k is the number of clusters. The value of k is determined experimentally as there is no prior knowledge about how many clusters are there. A too large k value would result in over-clustering and increase the number of false negatives, while a too small k value would not help much in reducing the search space. According to our experiment, in which there are 9,800 images with 82,552 regions, we divide the entire set of image regions into 100 clusters since it results in a good balance between accuracy and efficiency. d is some distance measure. In this study, we use the Euclidean distance.

2.2.3 Initialization

The initial size of population is set to l which is 50 in this study. For each chromosome we randomly generate k genes, which are actually k integers between 1 and n (the number of image regions). These k genes correspond to the representative image region for each of the k clusters. We then calculate the inverse values of the objective function for these chromosomes: f_1, f_2, ... , f_l. The fitness of each individual chromosome is computed according to Equation (2).

$$Fit_i = f_i / \sum_{i=1}^{l} f_i \tag{2}$$

2.2.4 Genetic Operators

1) *Selection*: There are many kinds of selection operations. We use a Roulette to simulate the selection as shown in Fig. 3.

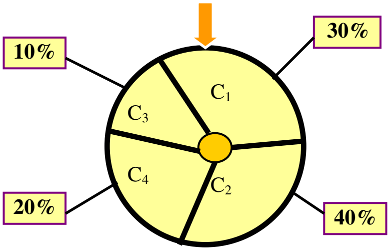


Fig. 3. Roulette

For each chromosome we compute its fitness according to Equation (2). Two chromosomes from the population are randomly selected. The higher the fitness the higher the chance a chromosome is selected. This mechanism is like rotating roulette as shown in Fig. 3. C_1, C_2... are chromosomes. The area each chromosome occupies is determined by its fitness. Therefore, chromosomes with higher fitness values would have more chances to be selected in each rotation. We select l pairs of chromosomes and feed them into the next step.

2) *Recombination*: In this step, the recombination operator proposed in [9] is used instead of a simple crossover. Given a pair of chromosomes C_1 and C_2, recombination operator generates their child C_0 one gene at a time. Each gene in C_0 is either in C_1 or C_2 or both and is not repetitive of other genes in C_0.

3) *Mutation*: In order to obtain high diversity of the genes, a “newly-born” child chromosome may mutate one of its genes to a random integer between 1 and n. However, this mutation is operated at a very low frequency.

2.2.5 Wrap-Up

At the end of clustering, we choose from the last generation the chromosome with the greatest fitness value. Thus we have k clusters. Given a query image region, all the other image regions in this cluster can be located. However, we cannot simply reduce the search space to this cluster because it is often the case that a particular region is closer to some regions in another cluster than some regions within the same cluster. This situation is illustrated in Fig. 4, where the query region A is closer to B than to C. Therefore we choose three clusters whose centroids are the closest to the query region. As an image is composed of several semantic regions, it can fall into any cluster that has at least one of its semantic regions. We then group all the images that have at least one semantic region fall into the three clusters and take it as the reduced search space for a given query region.

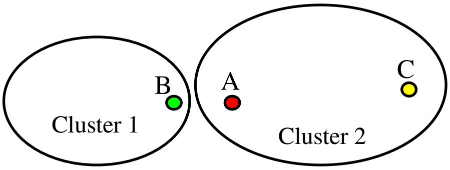


Fig. 4. Cluster Result

3 The Proposed Learning Approach

In this study, we assume that user is only interested in one semantic region of the query image. The goal is to retrieve those images that contain similar semantic regions. In the proposed CBIR system, we adopted the automatic image segmentation method proposed in Blobworld [3]. After segmentation, 8 global features (three texture features, three color features and two shape features [3]) for each "blob" i.e. semantic region, are extracted.

3.1 Multiple Instance Learning with Relevance Feedback

In traditional supervised learning, each object in the training set has a label. The task of learning is to map a given object to its label according to the information learned from the training set. However, in Multiple Instance Learning (MIL), the label of an individual instance is unknown. Only the label of a bag of instances is available. MIL needs to map an instance to its label according to the information learned from the bag labels. In CBIR, each image is considered a bag of semantic regions (instances). By supplying feedback to the retrieved images, user labels an image positive if it contains the region of interest; otherwise, it is labeled negative. As a result, the label of each retrieved image bag is available. However, the labels of the semantic regions are still unknown. The goal of MIL, in the context of CBIR, is to estimate the labels (similarity scores) of the test image regions/instances based on the learned information from the labeled images/bags. In this way, the single region based CBIR problem can be transformed to a MIL problem as defined below.

Definition 1. *Given a set of training examples T=<B,L> where $B=B_i(i=1,...,n)$ is a set of n bags and $L=L_i(i=1,...,n)$ is a set of labels of the corresponding bags. $L_i \in \{1$(Positive), 0(Negative)$\}$ The goal of MIL is to identify the label of a given instance in a given bag.*

The relation between a bag (image) label and the labels of all its instances (regions) is defined as below.

$$L_i = 1 \quad if \quad \exists_{j=1}^{m} l_{ij} = 1 \tag{3}$$

$$L_i = 0 \quad if \quad \forall_{j=1}^{m} l_{ij} = 0 \tag{4}$$

Suppose there are m instances in B_i. l_{ij} is the label of the j^{th} instance in the i^{th} bag. If the bag label is positive, there exists at least one positive instance in that bag. If the bag label is negative, all instances in that bag are negative. In this study, the One-Class SVM is adopted as the underlying learning algorithm.

3.2 One-Class SVM

One-Class classification is a kind of unsupervised learning. It tries to assess whether a test point is likely to belong to the distribution underlying the training data. In our case, the training set is composed of positive samples only. One-Class SVM has so far been studied in the context of SVMs [1].

The idea is to model the dense region as a "ball". In MIL problem, positive instances are inside the "ball" and negative instances are outside. If the origin of the "ball" is $\vec{\alpha}$ and the radius is r, a point $\vec{x_i}$, in this case an instance (image region) represented by an 8-feature vector, is inside the "ball" *iff* $\left\|\vec{x_i} - \vec{\alpha}\right\| \le r$. This is shown in Fig. 5 with red rectangles inside the circle being the positive instances.

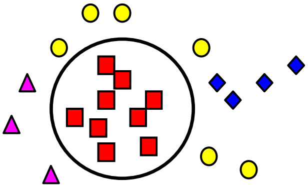


Fig. 5. One-Class Classification

This "ball" is actually a hyper-sphere. The goal is to keep this hyper-sphere as "pure" as possible and include most of the positive objects. Since this involves a non-linear distribution in the original space, the strategy of Schölkopf's One-Class SVM is first to do a mapping θ to transform the data into a feature space F corresponding to the kernel K:

$$\theta(u) \cdot \theta(v) \equiv K(u, v) \tag{5}$$

where u and v are two data points. In this study, we choose to use Radial Basis Function (RBF) Machine below.

$$K(u, v) = \exp\left(\left\|u - v\right\| / 2\sigma\right) \tag{6}$$

Mathematically, One-Class SVM solves the following quadratic problem:

$$\min_{w,\xi,\rho} \frac{1}{2}\|w\| - \alpha\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i \tag{7}$$

subject to

$$(w \cdot \theta(x_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0 \text{ and } i = 1,\ldots,n \tag{8}$$

where ξ_i is the slack variable, and $\alpha \in (0,1)$ is a parameter that controls the trade off between maximizing the distance from the origin and containing most of the data in the region created by the hyper-sphere and corresponds to the ratio of "outliers" in the training dataset. When it is applied to the MIL problem, Equation (7) is also subject to Equations (3) and (4). If w and ρ are a solution to this problem, then the decision function is $f(x) = sign(w \cdot \theta(x) - \rho)$ and it will be 1 for most examples x_i contained in the training set.

3.3 Learning and Retrieval Process

In initial query, user identifies a semantic region of his/her interest. We simply compute the Euclidean distances between the query semantic region and all the other semantic regions in the image database. The similarity score for each image is then set to the inverse of the minimum distance between its regions and the query region. The training sample set is then constructed according to user's feedback. If an image is labeled positive, its semantic region that is the least distant from the query region is labeled positive. For some images, Blob-world may "over-segment" such that one semantic region is segmented into two or more segments. In addition, some images may actually contain more than one positive region. Therefore, we cannot assume that only one region in each image is positive. Suppose the number of positive images is h and the number of all semantic regions in the training set is H. Then the ratio of "outliers" in the training set is set to:

$$\alpha = 1 - (\frac{h}{H} + z) \tag{9}$$

z is a small number used to adjust the α in order to alleviate the above mentioned problem. Our experiment results show that z= 0.01 is a reasonable value.

The training set as well as the parameter α are fed into One-Class SVM to obtain w and ρ, which are used to calculate the value of the decision function for the test data, i.e. all the image regions in the database. Each image region will be assigned a "score" by $w \cdot \theta(x) - \rho$ in the decision function. The similarity score of each image is then set to the highest score of all its regions. It is worth mentioning that except for the initial query in which the user needs to specify the query region in the query image, the subsequent iterations will only ask for user's feedback on the whole image.

4 Experiments

Fig. 6 shows the architecture of our system. Fig. 7 shows the initial query interface. The leftmost image is the query image. This image is segmented into 7 semantic regions (outlined by red lines). User identifies the "red flower" region as the region of interest (the 3rd image from left outlined by a blue rectangle). In initial query, the system gets the feature vector of the query region and compares it with those of other image regions using Euclidean distance. After that, user gives feedback to the retrieved images. Our One-Class SVM based algorithm learns from these feedbacks and starts another round of retrieval.

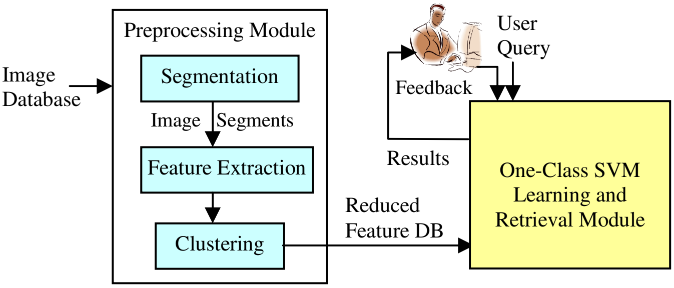


Fig. 6. CBIR System Architecture

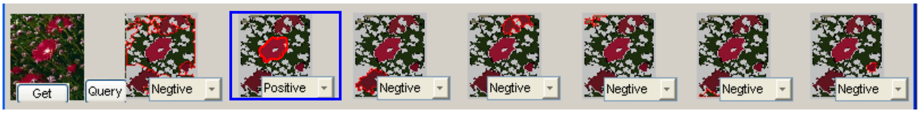


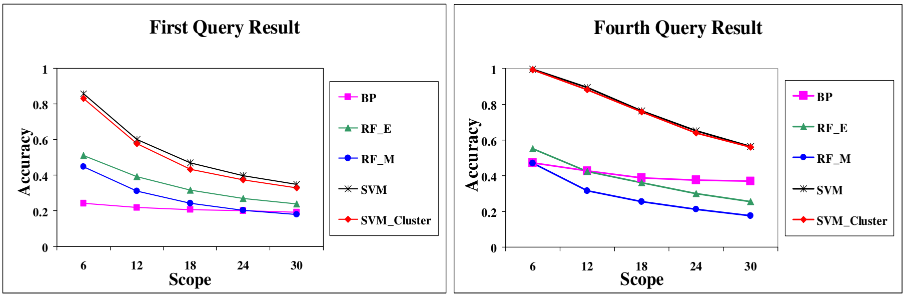
Fig. 7. Initial Query Interface

4.1 System Performance Evaluation

The experiment is conducted on a Corel image database consisting of 9,800 images. After segmentation, there are 82,552 image segments. Experiments show that when the number of clusters k=100, the result is most reasonable in terms of the balance between accuracy and reduction of search space. According to user-specified query region, we pull out three closest clusters as the reduced search space. Sixty five images are randomly chosen from 22 categories as the query images. The search space, in terms of the number of images in the 3 candidate clusters, is reduced to **28.6%** of the original search space on average.

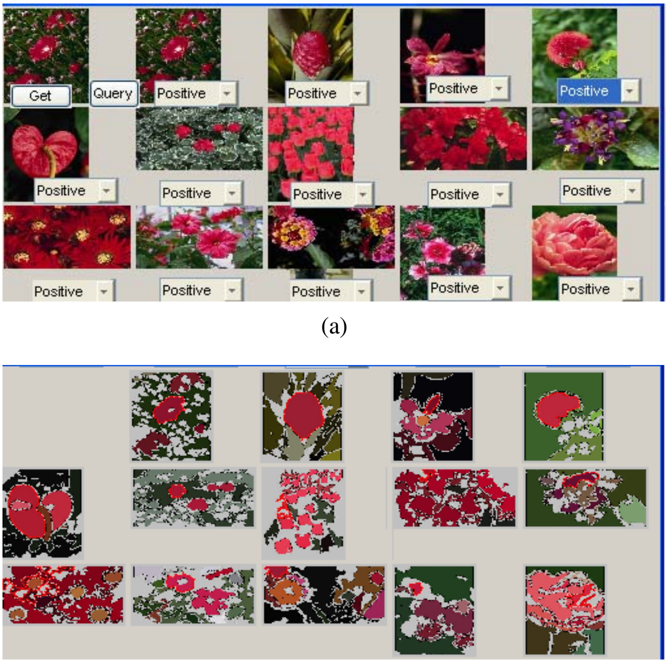
We compare our system with the one that performs full search. We also compare its performance with two other relevance feedback algorithms: 1) Neural Network based MIL algorithm [7]; 2) General feature re-weighting algorithm [2]. For the latter, both Euclidean and Manhattan distances are tested.

Five rounds of relevance feedback are performed for each query image - Initial (no feedback), First, Second, Third, and Fourth. The accuracy rates with different scopes, i.e. the percentage of positive images within the top 6, 12, 18, 24 and 30 retrieved images, are calculated. Fig. 8(a) shows the result from the First Query while Fig. 8(b) shows the result after the Fourth Query. "BP" is the Neural Network based MIL which uses both positive and negative examples. “RF_E” is feature re-weighting method with Euclidean Distance while “RF_M” uses Manhattan Distance. “SVM_Cluster” is the proposed system and “SVM” refers to the same retrieval mechanism without clustering.



(a) (b)

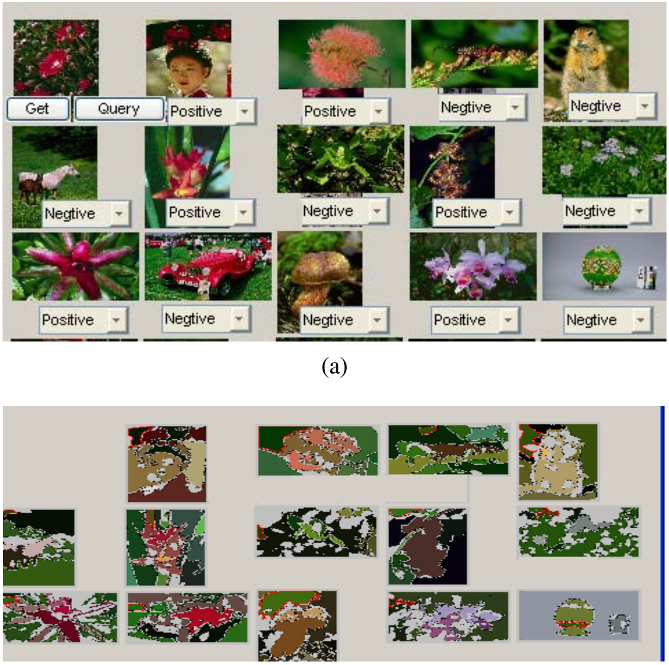
Fig. 8. (a) Retrieval Accuracy after the 1st Query; (b) Retrieval Accuracy after the 4th Query



(a)

(b)

Fig. 9. (a) Third Query Results by “SVM_Cluster”. **(b)** Retrieved Regions of the Images in Fig. 9(a)



(a)

(b)

Fig. 10. (a) Third Query Results by Neural Network based MIL**. (b)** Retrieved Regions of the Images in Fig. 10(a)

It can be seen from Fig. 8 that although the search space is substantially reduced, the performance of our system is only slightly worse than that of the ‘SVM’ without clustering. In addition, the accuracy of the proposed algorithm outperforms all other three algorithms.

We further compare “SVM_Cluster” with “BP” by examining the exact image regions learned by the two algorithms. Figures 9(a) and 10(a) show the Third Query results of “SVM_Cluster” and “BP”, respectively, given the query image as in Fig. 7. Figures 9(b) and 10(b) are the corresponding regions (outlined by red lines) learned by the two algorithms. It can be seen that, although “BP” seems to successfully find several “red flower” images, the regions it retrieved are actually the grass. Consequently, the “red flower” images in Fig. 10(a) will be labeled positive by the user. This will definitely affect the next round of learning. The bad performance of “BP” is due to excessive influence of “negative” samples.

5 Conclusion

In this paper, we proposed a MIL framework for single region based CBIR systems. In preprocessing, the search space is substantially reduced by using a clustering mechanism based on Genetic Algorithm. We then adopt One-Class SVM in the image retrieval phase. The advantage of our algorithm is that it targets image region retrieval instead of the whole image, which is more reasonable since the user is often interested in only one region in the image. The proposed work also transfers the One-Class SVM learning for region-based CBIR into a MIL problem. Due to the robustness of Genetic Algorithm in approximating global optima and the generality of One-Class SVM, the proposed system can better identify user's real need and remove the noise data.

Acknowledgement

The work of Chengcui Zhang was supported in part by SBE-0245090 and the UAB ADVANCE program of the Office for the Advancement of Women in Science and Engineering.

References

1. Schölkopf, B., Platt, J.C. et al: Estimating the Support of a High-dimensional Distribution. Microsoft Research Corporation Technical Report MSR-TR-99-87, 1999.
2. Rui, Y., Huang, T.S., and Mehrotra, S.: Content-based Image Retrieval with Relevance Feedback in MARS. Proc. of the Intl. Conf. on Image Processing, pp. 815-818, 1997.
3. Carson, C., Belongie, S., Greenspan, H., and Malik, J.: Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No.8, 2002.
4. Su, Z., Zhang, H.J., S. Li, and Ma, S.P.: Relevance Feedback in Content-based Image Retrieval: Bayesian Framework, Feature Subspaces, and Progressing Learning. IEEE Trans. on Image Processing, Vol. 12, No. 8, pp. 924-937, 2003.
5. Maron, O. and Lozano-Perez, T.: A Framework for Multiple Instance Learning. Advances in Natural Information Processing System 10. Cambridge, MA, MIT Press, 1998.
6. Chen, Y., Zhou, X., Tomas, S., and Huang, T.S.: One-Class SVM for Learning in Image Retrieval. Proc. of IEEE International Conf. on Image Processing, 2001.
7. Huang, X., Chen, S.-C., Shyu, M.-L., and Zhang, C.: User Concept Pattern Discovery Using Relevance Feedback and Multiple Instance Learning for Content-Based Image Retrieval. Proc. of the 3rd Intl. Workshop on Multimedia Data Mining (MDM/KDD'2002), pp. 100-108, 2002.
8. Holland, J. H.: Adaptation in Natural and Artificial Systems. University of Michigan Press (1975).
9. Vladimir, E. C. and Murray, A. T.: Spatial Clustering for Data Mining with Genetic Algorithms. Technical Report FIT-TR-97-10, Queensland University of Technology, Faculty of Information Management, September 1997.
10. Gondra, I. and Heisterkamp, D. R.: Adaptive and Efficient Image Retrieval with One-Class Support Vector Machines for Inter-Query Learning. WSEAS Transactions on Circuits and Systems, Vol. 3, No. 2, April 2004, pp. 324-329.