Image Classification Using Feature Subset Selection

SANG-SUNG PARK[1], KWANG-KYU SEO[2] , HO-SEOK MOON[1], YOUNG-GEUN SHIN[1], DONG-SIK JANG[1]

[1] Industrial Systems and Information Engineering, Korea University
1, 5-ka, Anam-Dong, Sungbuk-Ku, Seoul 136-701, KOREA
formation and Systems Engineering, Sangmyung University
San 98-20, Anso

Abstract: Classification technology is essential for fast retrieval in large database. This paper proposes a combining GA and SVM model to content-based image retrieval. The proposed method is also used to classification similar images from database. Joint HSV histogram and average entropy computed from gray-level co-occurrence matrices in the localized image region is employed as input vectors. Genetic algorithm is employed to select feature subsets eliminated irrelevant factors as used inputs and to determine the optimal parameters of Support Vector Machine. Experimental results show that the proposed model outperforms existing method.

Key-Words: SVM, Genetic Algorithm, CBIR, Feature Selection, Image

1 Introduction

The development of information technologies makes the demand of multimedia in-formation services significant. Recent research on retrieval methods has become very important for image and video searches.
In this paper, we deal with content-based image retrieval, which is a technique to retrieve images based on their visual properties such as color [1], texture [2], and shape [3, 4]. Systems [5, 6, 7] are well known for supporting this content-based image retrieval. Fast retrieval in databases has been one of the active research areas. In that process, without any clustering schemes and adequate indexing structures, retrievals of similar images are time-consuming because the database system must compare the query image to each image in the database. This cost can be particularly prohibitive if the database images are very large and their features tend to have high-dimensionality. This high-dimensional indexing structure increases the retrieval time and memory space exponentially, as the member of feature dimension increases. Thus, frequently, it does not have any advantages against the simple sequential search. So, fast search algorithms, which can deal with high-dimensional feature data, are often an essential component of the image database. There have been a number of indexing data structures suggested to handle high-dimensional data [8, 9, 10].

In order to classify images efficiently, we need to learn the previous image patterns. This can improve the accuracy of image classification and detection. In addition, we need to classify the images from a large and complex database. In this respect, we propose a new image classification technique based on SVM (Support Vector Ma-chine) that is useful for speedily finding the images from a large image database sys-tem. In this scheme, similar images are classified based on the image feature and associated classification algorithm. When the query is presented, similar images to the query are retrieved only from the most similar cluster to the query, thus full-database searches are not necessary.
We use a hybrid model with combining GA(Genetic Algorithm) and SVM as clustering technique for narrowing the search space. GAs are computational models of evolution. They work on the basis of a set of candidate solutions. The SVM is a training algorithm for learning classification and regression rules from data. In this study, GA is employed to select feature subsets eliminated irrelevant factors as used inputs and to determine the optimal parameters of SVM.

2 Image Features

In order to perform the content-based image retrieval, features which are representative of image content,

should be extracted. In this paper, color and texture information are used to represent image features. For color, joint HSV histogram extracted from local region is employed. For texture, entropies computed from local region are employed. These features extracted from each image in the database are used as input vector to the classifier.

Color: For representing color, we used HSV (Hue, Saturation, Value) color model because this model is closely related to human visual perception. For color quantization, we uniformly quantized HSV space into 18 bins for hue (each bin consisting of a range of 20 degree), 3 bins for saturation and 3 bins for value for lower resolution.
In order to represent the local color histogram, we divided image into equal-sized rectangular regions and extract HSV joint histogram that has quantized 162 bins for each region. And to obtain compact representation, we extract from each joint histogram the bin that has the maximum peak. The HSV representation of an image from RGB is obtained using the following relationships:

$$H = \begin{cases} \theta, & G \geq B, \\ 2\pi - \theta & G \leq B, \end{cases} \qquad (1)$$

$$where \quad \theta = \cos^{-1}\left[\frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)^2+(R-B)(G-B)]^{\frac{1}{2}}} \right],$$

$$S = 1 - \frac{3}{R+G+B}\left[\min(R,G,B)\right], \qquad (2)$$

$$V = \frac{1}{3}(R+G+B). \qquad (3)$$

Take hue , saturation , and value associated to the bin as representing features in that rectangular region and normalize to be within the same range of [0,1]. Thus, each image has the dimensional color vector.

Texture: Most natural images include textures. Scenes containing pictures of wood, grass, etc. can be easily classified based on the texture rather than color or shape. Therefore, it may be useful to extract texture features for image clustering. Like as color feature, we include a texture feature extracted from localized image region.

As a texture feature, we used the entropy extracted from the co-occurrence matrix [5]. Detailed feature extraction is performed as follows:

1. Conversion of color image to gray image
2. Dividing image into 3×3 rectangular regions as in color case.
3. Obtaining co-occurrence matrix for four (horizontal 0^0, vertical 90^0 and two diagonal 45^0 and 135^0) orientation in region and normalize entries of four matrices to [0, 1] by dividing each entry by total number of pixels.
4. Extracting average entropy value from four matrices.

$$e = \frac{-\sum_k \sum_i \sum_j p(i,j)\log(p(i,j))}{4}, \qquad k = 1,2,3,4 \qquad (4)$$

5. Constructing texture feature vector by concatenating entropies over all rectangular regions.

Thus, each image has the 3×3(=9) dimensional texture vector.

3 GA(Genetic Algorithm)

GAs are computational models of evolution. They work on the basis of a set of candidate solutions. Each candidate solution is called a 'chromosome', and the whole set of solutions is called a 'population'. The algorithm allows movement from one population of chromosomes to a new population in an iterative fashion. Each iteration is called a 'generation'. There are various forms of GAs, a simple version, which is called static population model was used in all the experiments [11, 12].
In the static population model, the population is ranked according to the fitness value of each chromosome. At each generation, two (and only two) chromosomes are selected as parents for reproduction. GAs operate iteratively on a population of structures, each one of which represents a candidate solution to the problem at hand, properly encoded as a string of symbols (e.g.,binary). A randomly generated set of such strings forms the initial population from which the GA starts its search. Three basic genetic operators guide this search: selection, crossover, and mutation. The genetic search process is iterative: evaluating, selecting, and recombining strings in the population during each iteration (generation) until reaching some termination condition. The basic algorithm, where P(t)

is the population of strings at generation t, is given below:

```
t = 0
initialize P(t)
evaluate P(t)
while (termination condition is not satisfied) do
begin
        select P(t + 1) from P(t)
                recombine P(t + 1)
                evaluate P(t + 1)
                t = t + 1
end
```

Evaluation of each string is based on a 1tness function that is problem-dependent. It determines which of the candidate solutions are better. This corresponds to the environmental determination of survivability in natural selection. Selection of a string, which represents a point in the search space, depends on the string's 1tness relative to those of other strings in the population. It probabilistically removes, from the population, those points that have relatively low fitness. Mutation, as in natural systems, is a very low probability operator and just flips a specific bit. Mutation plays the role of restoring lost genetic material. Crossover in contrast is applied with high probability. It is a randomized yet structured operator that allows information exchange between points. Its goal is to preserve the fittest individuals without introducing any new value.

4 SVM(Support Vector Machine)

The support vector machine (SVM) [13, 14, 15] is a training algorithm for learning classification and regression rules from data, for example the SVM can be used to learn polynomial, radial basis function (RBF) and multi-layer perceptron (MLP) classifiers . SVMs were first suggested by Vapnik in the 1960s for classification and have recently become an area of intense research owing to developments in the techniques and theory coupled with extensions to regression and density estimation.

SVMs arose from statistical learning theory; the aim being to solve only the problem of interest without solving a more difficult problem as an intermediate step. SVMs are based on the structural risk minimization principle, closely related to regularization theory. This principle incorporates capacity control to prevent over-fitting and thus is a partial solution to the bias-variance trade-off dilemma. Two key elements in the implementation of SVM are the techniques of mathematical programming and kernel functions. The parameters are found by solving a quadratic programming problem with linear equality and inequality constraints; rather than by solving a non-convex, unconstrained optimization problem. The flexibility of kernel functions allows the SVM to search a wide variety of hypothesis spaces. For constructing the decision rules, four common types of SVM are given as follows:

- Linear: $K(x_i, x_j) = x_i^T x_j$ (5)

- Polynomial: $K(x_i, x_j) = (x_i^T x_j + r)^d$ (6)

- Radial basis function (RBF):

$$K(x_i, x_j) = \exp(-\left\|x_i - x_j\right\|^{2/2\delta^2}) \quad (7)$$

- Sigmoid: $K(x_i, x_j) = \tanh(x_i^T x_j + r)$ (8)

5 Proposed Algorithm

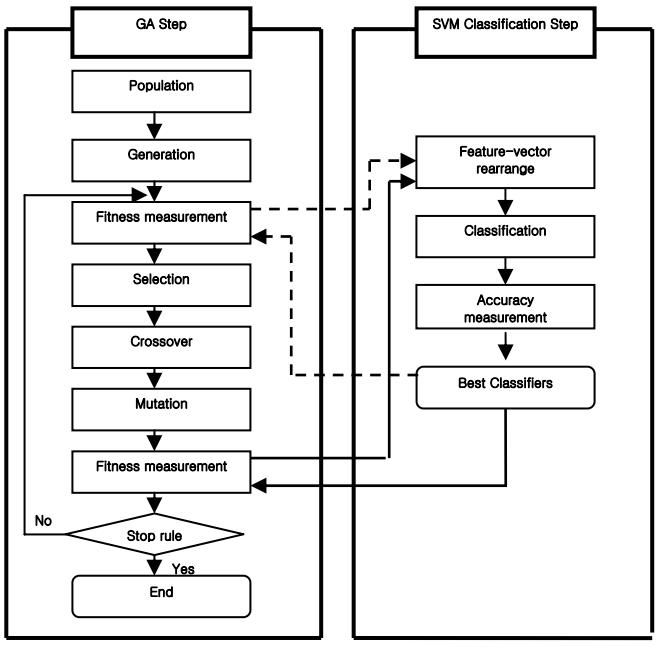

Fig 1. The flow chart of proposed algorithm

This paper proposes a hybrid model with combining GA and SVM. In this study, GA is employed to select feature subsets eliminated irrelevant factors as used inputs and to determine the optimal parameters of SVM. The flow chart of the proposed algorithm is depicted in Fig.1. The proposed algorithm in Fig 1 is to optimize SVM's variables and input data for Image retrieval. The procedure of proposed algorithm begins by selecting random chromosome in the population which is represented by string input data and SVM variables. Each strings will sent to SVM classifier and evaluate the fitness. The SVM model is used to obtain hit ration of each chromosome. The fitness in this study is given below:

$$F = \lambda_1 Q_1 + \lambda_2 \frac{1}{Q_2} \tag{9}$$

Q_1 is accuracy of each class which is classified by using subset. $Q2$ is number of selected eigenvector. We define λ_1 is 100 and λ_2 is 10 in the experiment.

6 Experiments

To show the effective classification of the proposed method, we checked the classification accuracy. All experiments were performed on a Pentium IV with 512 Mbytes of main memory and 100Gbytes of storage. All programs have been implemented in Visual C++. We experimented on 1,200 images where most of them have dimensions of 192×128 pixels. The 1,200 images can be divided into 6 categories each with 200 images such as horse, rose, polar bear, sunset, valley and dolphin.

We performed two experiments:

1) Classification results according to kernel of different types.

2) Classification results using SVM classifier and proposed classifier.

As shown in Table 1, both training and test success rates that were achieved under each different method. As can be seen, proposed classification with RBF kernel has consistently given the best performance of other. The average classification of 6 classes with RBF kernel achieves 96.93% success on the training set and 96.17 % with the test set.

Image	Type of kernel	Training (%)	Testing (%)
Horse	Linear	97.1	94.0
	Polynomial	99.0	97.8
	RBF	100	99.5
	Sigmoid	99.5	98.8
Rose	Linear	90.1	89.3
	Polynomial	93.8	92.7
	RBF	96.8	95.9
	Sigmoid	95.2	94.7
Polar - Bear	Linear	97.4	95.7
	Polynomial	99.1	97.3
	RBF	100	99.0
	Sigmoid	100	98.6
Sunset	Linear	96.8	96.0
	Polynomial	98.7	96.6
	RBF	100	99.4
	Sigmoid	97.6	96.9
Valley	Linear	82.4	79.0
	Polynomial	84.9	81.9
	RBF	87.4	87.0
	Sigmoid	84.8	80.8
Dolphin	Linear	92.2	90.6
	Polynomial	97.2	93.2
	RBF	97.4	96.2
	Sigmoid	97.3	94.3
Average	Linear	92.67	90.77
	Polynomial	95.45	93.25
	RBF	96.93	96.17
	Sigmoid	95.73	94.02

Table 1. The performance of proposed classification according to kernel type

Table 2 shows classification results using SVM classifier and proposed classifier. SVM classification shows average accuracy 91.65%, whereas proposed classification shows average accuracy 93.87%.

Image	SVM(%)	GA+ SVM(%)
Horse	92.7	99.5
Rose	94.1	95.9
Polar Bear	95.7	99
Sunset	96.0	99.4
Valley	80.8	87
Dolphin	90.6	96.2
Average	91.65	95.87

Table 2. Classification results using SVM classification and proposed classification

7 Conclusion

In this paper proposes a combining GA and SVM model for content-based image retrieval. As input elements to system, dominant triple (hue, saturation, and value) which are extracted from quantized HSV joint histogram are used for representing color information and average entropy computed gray-level co-occurrence matrices are used for texture information in the image. The proposed method served to exemplify that kernel-based learning algorithms are indeed highly competitive on variety problems with different characteristics and can be employed as an efficient method for CBIR.

The study needs further research as follows. The selection of the kernel function and the determination of optimal values of the parameters have a critical impact on the performance in SVM. Therefore it is necessary to investigate to develop a structured method of selecting an optimal value of parameters and kernel function in SVM for the best prediction performance. In addition, we develop the generalization of SVM on the basis of the appropriate level of the training set size and give a guideline to measure the generalization performance.

Acknowledgement:

This work was supported by the Brain Korea 21 Project in 2006.

References:

[1] Smith, J.R., Chang, S.F., Tools and techniques for color image retrieval, *In Proc. SPIE: Storage and Retrieval for Image and Video Databases IV*, Vol. 2670, 1996, pp.426-437.

[2] Manjunath, B.S., Ma, W.Y., *Texture features for browsing and retrieval of image data*, Tech. Rep. CIPR TR, 95-06, 1995.

[3] Jain, A.K., Vailaya, A., Shape-based retrieval: A case study with trademark image databases. *Pattern Recognition*, Vol. 31, No. 9, 1998, pp. 1369-1390.

[4] Swain, M., Ballard, D., Color indexing, *International Journal of Computer Vision*, Vol. 7, No. 1., 1991, pp.11-32.

[5]Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafer, J., Lee, D., Petkovic, D., Steele, D., Yanker, P., Query by image content: The QBIC system, *IEEE Computer*, Vol. 28, No. 9., 1995, pp.23-31.

[6] Smith, J.R., Chang, S.E., VisualSEEK: A fully automated content-based image query system, *In Proc. ACM Multimedia*, 1996, pp.87-98

[7] Carson, C., Belongie, S., Greenspan, H., Malick, 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, pp. 1026-1638.

[8] White, D.A., Jain, R., Similarity indexing with the SS-tree, *In Proc. 12th IEEE Inter-national Conference on Data Engineering*, 1996, pp. 516-523.

[9] Lin, K.I., Jagadish, H.V., Faloutsos, C., The TV-tree: An index structure for high-dimensional data, *VLDB Journal*, Vol. 3, No. 4., 1994, pp. 517-549.

[10]Berchtold, S., Keim, D.A., Kriegel, H.P., The X-tree: An index structure for high-dimensional data, *In Proc. 22th Int. Conf. on Very Large Data Bases*, 1996, pp.28-39.

[11]D. Whiney, *A genetic algorithm tutorial. Technical Report*, Department of computer science, Colorado state university, 1993, CS-93-103

[12]R. L. Haupt, An introduction to genetic algorithms for electromagnetics, *IEEE Magazine, Antennas Propagation*, Vol. 37., 1995, pp.7-15.

[13]V. Vapnik., *Statistical Learning Theory*, Springer, New York, 1998.

[14]H. Drucker, D. Wu, and V.N. Vapnik., Support vector machines for spam catergori-zation , *IEEE Transactions on Neural Networks*, Vol. 10, No. 5., 1999, pp.1048-1054.

[15]A. Fan and M. Palaniswami., Selecting bankruptcy predictors using a support vector machine approach, *Proceedings of the International Joint Conference on Neural Networks*, 2000.