GA-SVM and Mutual Information based Frequency Feature Selection for Face Recognition

AOUATIF AMINE[1]
ALI EL AKADI[2]
MOHAMMED RZIZA [3]
DRISS ABOUTAJDINE[3]

GSCM-LRIT, Faculty of Sciences,
Mohammed V University,
B.P. 1014 Rabat, Morocco
[1]aouatif.amine@ieee.org
[2]elakadi@yahoo.com
[3]{rziza,aboutaj}@fsr.ac.ma

Abstract. The dimensionality of existing data make it difficult to deploy any information to identify features that discriminate between the classes of interest. Feature selection involves reducing the number of features, removes irrelevant, noisy and redundant data without significantly decreasing the prediction accuracy of the classifier. An efficient feature selection and classification technique for face recognition is presented in this paper. Genetic Algorithms (GAs) for feature selection and Support Vector Machine (SVM) for classification are incorporated in the proposed technique. The proposed GAs-SVM technique has two purposes in this research: Selecting of the optimal feature subset and Selecting of the kernel parameters for SVM classifier. The input feature vector for the GAs-SVM are extracted by using the Discrete Cosine Transform (DCT). We evaluate its efficiency compared to the recently proposed feature selection algorithm based on mutual information. The results show that the proposed approach is promising, it is able to select small subsets and still improve the classification accuracy.

Keywords: Face recognition, Feature Selection, Mutual Information, Genetic Algorithm, Support Vector Machine, Discrete Cosine Transform.

(Received August 29, 2008 / Accepted January 19, 2009)

1 Introduction

Machine recognition of faces is becoming more and more popular and the need for accurate and robust performance is increasing. Face recognition, as an unsolved problem under the conditions of pose and illumination variations, still attracts significant research efforts. The main reasons for the ongoing research are: (i) the increased need for natural identification for authentication in the networked society, for surveillance, for perceptual user interfaces, (ii) and the lack of robust features and classification schemes for the face recognition task.

A common objective in face recognition is to find a good way of representing face information. High information redundancy present in face images results in inefficiencies when these images are used directly for recognition, identification and classification. A key point in developing a good representation is to expose the constraints and remove the redundancies contained in pixel images of faces. Typically one builds a computational model to transform pixel images into face features, which generally should be robust to variations of illumination, scale and orientation and then use these features for recognition [12]. For classical pattern recognition techniques, the patterns are generally represented as a vector of feature values. The selection of features can have a considerable impact on the effectiveness of the resulting classification algorithm [27]. It is not often known in advance wich features will provide the best discrimination between classes, and it is usually not feasible to measure and represent all possible features of the objects effects. With feature selection, the cost of classification can be reduced by limiting the number of features which must be measured and stored. Some, but not all, feature selection methods realize this benefit as well.

A number of approaches for feature subset selection have been proposed in the literature. Koller et al [30] used a greedy algorithm to remove the features that provide the least additional information given the remaining features. Brill et. al [9] have explored randomized population-based heuristic search approaches such as GAs to select feature subsets for NNs. As is known, in many supervised learning problems, feature selection is important, and for SVM, it also performs badly when there are many irrelevant features [36]. In order to improve its performance, suitable feature selection algorithm, such as MLR (Multiple Linear Regression), GA, should be adopted. GAs are good candidates for attacking this challenge since GAs are very useful for extracting patterns in multiclass [20], high-dimensionality problems where heuristic knowledge is sparse or incomplete.

Feature selection involves finding a subset of features to improve prediction accuracy or decrease the size of the structure without significantly decreasing the prediction accuracy of the classifier built by using only the selected features [1, 30]. There are many potential benefits of feature selection such as facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance [24].

Methods for feature selection are generally divided into three categories: the filter approach, the wrapper approach and the embedded method. In the first category, the filter approach is first utilized to select the subsets of features before the actual model learning algorithm is applied. On the other hand, the wrapper approach [29] utilize the learning machine as a fitness function and search for the best subset of features in the space of all feature subsets. Besides wrappers and filters, the embedded methods [24] are another category of feature selection algorithms, which perform feature selection in the process of training and are usually specific to given learning machines.

All feature selection methods needs to use an evaluation function together with a search strategy to obtain the optimal feature set. The evaluation function tries to measure the discriminating ability of a feature or a subset to distinguish the different class labels and can be grouped into five categories [16]: *distance, information (*or *uncertainty), dependence, consistency and classifier error*. Searching for the optimal subset can be achieved by examining all possible subsets, is usually unfeasible in practice due to the large amount of computational effort required. A wide range of heuristic search strategies have been used including forward selection [6], backward elimination [7], hill-climbing [11], branch and bound algorithms [35], and the stochastic algorithms like simulated annealing [17] and genetic algorithms (GAs) [23].

In this paper, we compared the effectivness of two feature selection approaches. Our aim is to study a dependency between a selected feature vector and the resulting accuracy. In order to see the relation between these parameters, we first use DCT to transform each image as a feature vector named Frequency Feature Subset (FFS). The two feature selection approaches are then used to select a subset of features from the low-dimensional representation by removing certain DCT coefficients that do not seem to encode important information about recognition task.

The first feature selection approach is a filter approach with four information-theoretic measures, and the second is new proposal wrapper using genetic algorithms and support vector machine classifier. For the classification process we used the SVM technique, which has proven to be efficient for nonlinearly separable input data, and in order to improve the SVM classification accuracy we implement a GA to automatize the choice of SVM parameters.

The rest of the paper is organized like follows: Some information theoretic notions for feature extraction and feature selection and genetic algorithm are addressed to the section 2. Section 3 is reserved to the SVM classifier for face recognition and the proposed GA-SVM technique for feature selection and parameters optimisation. The next section outlines the overview of the proposed method. Section 6 is dedicated to evaluate the performances of these methods in the context of face recognition problem. The last section summarizes the results and draws a general conclusion.

2 Theoretic Background

2.1 Discrete Cosine Transform

High information redundancy and correlation in face images result in inefficiencies when such images are used directly for recognition. DCT is a predominant tool first introduced by Ahmed et al. [2]. Since then, it was widely used as a feature extraction and compression in various applications on signal and image processing and analysis due to its fine properties, i.e., decorrelation, energy compaction, separability, symmetry and orthogonality. The 2-D DCT is a direct extension of the 1-D case and is given by:

$$C(u,v) = \frac{2}{M} \cdot \alpha(u)\alpha(v) \sum_{x=0}^{M-1} \sum_{y=0}^{M-1} I(x,y) \cdot \cos[\frac{(2x+1)u\pi}{2M}] \cos[\frac{(2y+1)v\pi}{2M}] \quad (1)$$

for $u, v = 0, 1, \ldots, M-1$

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{M}} & \text{for } u = 0; \\ 1 & \text{otherwise.} \end{cases}$$

For an $M \times N$ image, we have $M \times N$ DCT coefficient matrix covering all the spatial frequency components of the image. The DCT coefficients with large magnitude are mainly located in the upper-left corner of the DCT matrix. Accordingly, we scan the DCT coefficient matrix in a zigzag manner starting from the upper-left corner and subsequently convert it to a one-dimensional (1-D) vector. Detailed discussions about image reconstruction errors using only a few significant DCT coefficients can be found in [34].

In face recognition, DCTs are used to reduce image information redundancy because only a subset of the transform coefficients are necessary to preserve the most important facial features [25, 38]. In our study, we have used DCT for feature extraction.

2.2 Mutual information based measure for feature selection

2.2.1 Definitions and measurements

The first goal of a prediction model is to minimize the uncertainty on the dependent variable. A good formalization of the uncertainty of a random variable is given by Shannon and Weaver's [33] information theory. While first developed for binary variables, it has been extended to continuous variables. Let X and Y be two random variables (they can have real or vector values). We denote $\mu_{X,Y}$ the joint probability density function of X and Y. We recall that the marginal density functions are given by:

$$\mu_X(x) = \int \mu_{X,Y}(x,y)dy \quad (2)$$

$$\mu_Y(y) = \int \mu_{X,Y}(x,y)dx \quad (3)$$

Let us now recall some elements of information theory. The uncertainty on Y is given by its entropy defined as:

$$H(Y) = -\int \mu_Y(y) \log \mu_Y(y) dy \quad (4)$$

If we get knowledge on Y indirectly by knowing X, the resulting uncertainty on Y knowing X is given by its conditional entropy, that is:

$$H(Y|X) = -\int \mu_X(x) \int \mu_Y(y|X=x) \log \mu_Y(y|X=x) dydx \quad (5)$$

The joint uncertainty of the (X,Y) pair is given by the joint entropy, defined as:

$$H(X,Y) = -\int\int \mu_{X,Y}(x,y) \log \mu_{X,Y}(x,y) dxdy \quad (6)$$

The mutual information between X and Y can be considered as a measure of the amount of knowledge on Y provided by X (or conversely on the amount of knowledge on X provided by Y). Therefore, it can be defined as [15]:

$$I(X,Y) = H(Y) - H(Y|X) \quad (7)$$

which is exactly the reduction of the uncertainty of Y when X is known. If Y is the dependant variable in a prediction context, the mutual information is thus particularly suited to measure the pertinence of X in a model for Y [39]. Using the properties of the entropy, the mutual information can be rewritten into:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (8)$$

that is, according to the previously recalled definitions, into [13]:

$$I(X,Y) = -\int\int \mu_{X,Y}(x,y) \log \frac{\mu_{X,Y}(x,y)}{\mu_X(x)\mu_Y(y)} dxdy \quad (9)$$

The conditional mutual information is defined as:

$$I(X,Y|Z) = H(X|Y) - H(X|Y,Z) = I(X|Y,Z) - I(X|Y) \tag{10}$$

This value quantifies how much information is shared between X and Y, given the value of Z. Another way to see it, as it is decomposed above, is as the difference between the information required to describe X given Z, and the information to describe X given both Z and Y . If Y and Z carry the same information about X, the two terms on the right are equal, and the conditional mutual information is zero. On the opposite, if both Y and Z bring information, and if those informations are complementary, the difference is large.

Mutual information or information gain can be regarded as a measure of the strength of a 2-way interaction between an attribute X and the class Y. In this spirit, we can generalize it to 3-way interactions by introducing the interaction gain [28]:

$$I(X;Z;Y) = I(X,Z;Y) - I(X;Y) - I(Z;Y) \tag{11}$$

Interaction gain is identical to the notion of mutual information among three random variables.

2.2.2 Mutual Information Algorithms

Mutual information is a good indicator of relevance between variables, and have been used as a measure in several feature selection algorithms. The following sections will sketch four state-of-the-art filter approaches that use this quantity for feature selection.

Variable Ranking (Rank)

The ranking method returns a ranking of variables on the basis of their individual mutual information with the output. This means that, given n input variables, the method first computes n times the quantity $I(X_i, Y), i = 1, \ldots, n$, then ranks the variables according to this quantity and eventually discards the least relevant ones [18].

The main advantage of the method is its rapidity of execution. Indeed, only n computations of mutual information are required for a resulting complexity $O(n * 2 * N)$. The main drawback derives from the fact that possible redundancies between variables is not taken into account. Indeed, two redundant variables, yet highly relevant taken individually, will be both well ranked.

Minimum Redundancy - Maximum Relevance criterion (mRMR)

The minimum redundancy - maximum relevance criterion [31] consists in selecting the subset of feature X_S that maximizes the relevance term defined by:

$$D(X_S, Y) = \frac{1}{|S|} \sum_{X_i \in X_S} I(X_i; Y) \tag{12}$$

and minimize the redundancy term defined by:

$$R(X_S) = \frac{1}{|S|} \sum_{X_i, X_j \in X_S} I(X_i; X_j) \tag{13}$$

The mRMR feature set is obtained by optimizing the conditions in Eqs. 12 and 13 simultaneously. Optimization of both conditions requires combining them into a single criterion function. The two ways to combine relevance and redundancy, lead to two selection criteria:

(1) **mRMR-D**: Mutual Information Difference criterion:

$$\max(D(X_S, Y) - R(X_S)) \tag{14}$$

(2) **mRMR-Q**: Mutual Information Quotient criterion,

$$\max(D(X_S, Y)/R(X_S)) \tag{15}$$

In practice, incremental search methods can be used to find the near-optimal feature defined by Eqs. 14 and 15. More precisely, mRMR consists in selecting the variable X_i among the not yet selected features X_{-S} that maximizes the criterion below:

$$X_{mRMR-D} = \arg \max_{X_i \in X_{-S}} (I(X_i; Y) - \frac{1}{d} \sum_{X_j \in X_S} I(X_i; X_j)) \tag{16}$$

$$X_{mRMR-Q} = \arg \max_{X_i \in X_{-S}} \left[\frac{I(X_i; Y)}{\frac{1}{d} \sum_{X_j \in X_S} I(X_i; X_j))} \right] \tag{17}$$

Conditional Mutual Information Maximization Criterion (CMIM)

This approach [21] proposes to select the feature $X_i \in X_{-S}$ whose minimal conditional relevance $I(X_i; Y|X_j)$ among the selected features $X_j \in X_S$, is maximal. This requires the computation of the mutual information of X_i and the output Y, conditional on each feature $X_j \in X_S$ previously selected. Then, the minimal value is retained and the feature that has a maximal minimal conditional relevance is selected. the variable returned according to the CMIM criterion is:

$$X_{CMIM} = \arg \max_{X_i \in X_{-S}} (\min_{X_j \in X_S} (I(X_i; Y|X_j)) \quad (18)$$

Interaction Gain based Feature Selection (IGFS)

This criterion [3] is based on the individual Mutual Information and a compromise between features redundancy and features interaction. The compromise is made by the mean of Interaction Gain. In formal notation, the variable returned according to the IGFS criterion is:

$$X_{IGFS} = arg \underset{X \in X_{-S}}{max} (I(X_i; Y) + \frac{1}{d} \sum_{X_j \in X_S} I(X_i; X_j; Y)) \quad (19)$$

The main advantage in using this criterion for selecting variables is that an interacting variable of an already selected one has a much higher probability to be selected than with other criteria.

2.3 Genetic Algorithm (GA) based feature selection and parameters optimization

Genetic Algorithms (GAs) were developed by Holland in 1970. GAs are stochastic search algorithm modeled on the process of natural selection, which underlies biological evolution. GAs have been successfully applied in many search, optimization, and machine learning problems [23, 26]. GAs improve their ability to efficiently search large spaces about which little is known a priori. GA evolves a population of chromosomes as potential solutions to an optimization problem.

There are three major design decisions to consider when implementing a GA to solve a particular problem. A representation for candidate solutions must be chosen and encoded on the GA chromosome, fitness function must be specified to evaluate the quality of each candidate solution, and finally the GA run parameters must be specified, including which genetic operators to use, such as crossover, mutation, selection, and their possibilities of occurrence.

Initial Population

In general, the initial population is generated randomly. In this way, however, we will end up with a population where each individual contains the same number of $1's$ and $0's$ on the average. To explore subsets of different numbers of features, the number of $1's$ for each individual is generated randomly. Then, the $1's$ are randomly scattered in the chromosome.

Mutation

Mutation is the genetic operator responsible for maintaining diversity in the population (see Figure 1). Mutation operates by randomly "*flipping*" bits of the chromosome, based on some probability. A usual mutation probability is $1/p$, where p is the length of each of the two parts of the chromosomes. This probability should usually be set fairly low. If it is set to high, the search will turn into a primitive random search.

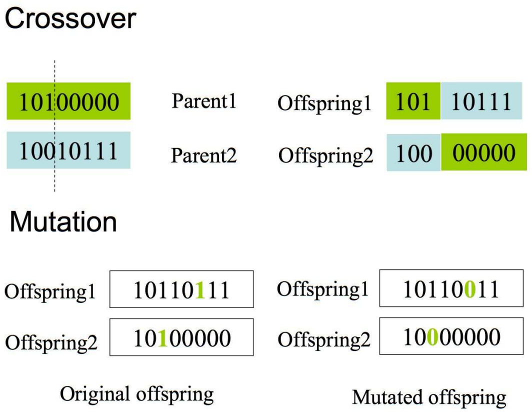

Figure 1: Genetic Crossover and Mutation process

Crossover

Crossover, the critical genetic operator that allows new solution regions in the search space to be explored, is a random mechanism for exchanging genes between two chromosomes using the one point crossover, two point crossover, or homologue crossover. Offspring replaces the old population using the elitism or diversity replacement strategy and forms a new population in the next generation (see Figure 1).

Replacement

Replacement schemes determine how a new population is generated. We used the concept of overlapping populations, where parents and offspring are merged, and the best individuals from this union will form the next population.

Selection

This is the process of choosing parents for reproduction. Usually, it emphasizes the best solutions in the population, but since the replacement scheme employed here already offers enough evolutionary pressure, a random selection approach was chosen.

Random immigrant

This is a method that helps to keep diversity in the population, minimizing the risk of premature convergence [14]. works by replacing the individuals whose fitness is under the mean by recently initialized individuals. Random immigrant is invoked when the best individual does not change for a certain number of generations.

Fitness Function

The main goal of feature selection is to use fewer features to obtain the same or better performance. Fitness function is one of the most important part in genetic search. This function (see Eq. 20) have to evaluate the effectivness of each individual in a population, so it has an individual as an input and it returns a numerical evaluation that must represent the goodness of the feature subset. The search strategy's goal is to find a feature subset maximizing this function. The crossover and mutation functions are the main operators that randomly impact the fitness value.

$$Fitness = SVM\_accuracy \tag{20}$$

3 SVM classifier for face recognition

3.1 Basic theory

Recently, the SVM has been gaining popularity in the field of pattern classification. SVM integrated pattern classification algorithm with non-linear formulation. It has the benifit that it can handle the classes with complex non-linear decision boundaries. SVM are binary classifiers and different approaches like "one-against-all" and "one-against-one" are built to extend SVM to the multi-class classification case for face recognition [10]. The major method is the "one-against-one" method. This method constructs classifiers where each one is trained on data from two classes. For training data from the i^{th} and the j^{th} classes, we solve the following binary classification problem:

$$\begin{aligned} \min_{w^{ij},b^{ij},\xi^{ij}} \frac{1}{2}(w^{ij})^T w^{ij} + C\sum_t \xi_t^{ij}(w^{ij})^T \\ (w^{ij})^T\phi(x_t) + b^{ij}1 - \xi_t^{ij}, if y_t = i \\ (w^{ij})^T\phi(x_t) + b^{ij}1 - \xi_t^{ij}, if y_t = i \\ \xi_t^{ij} \geq 0 \end{aligned} \tag{21}$$

There are different methods for doing the future testing after all $p(p-1)/2$ classifiers are constructed. After some tests, we decide to use the following voting strategy suggested in [22]: if $sign((w^{ij})^T\phi(x_t) + b^{ij})$ says x is in the i^{th} class, then the vote for the i^{th} class is added by one. Otherwise, the j^{th} is increased by one. Then we predict x is in the class with the largest vote. The voting approach described above is also called the "Max Wins" strategy. In case that two classes have identical votes, thought it may not be a good strategy, now we simply select the one with the smaller index. Practically we solve the dual of (Eq. 21) whose number of variables is the same as the number of data in two classes. Hence if in average each class has l/k data points, we have to solve $k(k-1)/2$ quadratic programming problems where each of them has about $2l/k$ variables.

3.2 Genetic algorithm for SVM parameters optimization

In the literature, only a few algorithms have been proposed for SVM feature selection like in [8]. Some other GA-based feature selection methods were proposed [32, 37]. However, these papers focused on feature selection and did not deal with parameters optimization for the SVM classifier. Therefore, in addition to the feature selection, proper parameters setting can improve the SVM classification accuracy. The choice of C and the kernel parameter is important to get a good classification rate. In the most case these parameters are tuned manually. In order to automatize this choice we use genetic algorithms. The SVM parameters, C and γ are real, we have to encode them with binary chains; we fix two search intervals, one for each parameter, $C_{max} \leq C \leq C_{min}$ and $\gamma_{max} \leq \gamma \leq \gamma_{min}$. To encode C and γ, we discretize the search spaces. Thus, a 32 bits encoding scheme of C is given by $C_{b1}, \ldots C_{b32}$ where:

$$C_b = \sum_{i=1}^{32} C_{bi}2^{i-1} \tag{22}$$

and γ by $\gamma_{b1}, \ldots \gamma_{b32}$ where:

$$\gamma_b = \sum_{i=1}^{32} \gamma_{bi}2^{i-1} \tag{23}$$

with $C_b = g_{max}(C - C_{min})/(C_{max} - C_{min})$ and $\gamma_b = g_{max}(\gamma - \gamma_{min})/(\gamma_{max} - \gamma_{min})$ and $g_{max} = 2^{32} - 1$.

The fitness function used to evolve the chromosomes population is the SVM classification rate. The goal was to see if the GA would discover the work effectively. We lists some reasons why SVM must be used combined feature selection. One major weakness of SVMs is their high computational cost, which precludes real-time applications. In addition, SVMs are formulated as a quadratic programming problem and, therefore, it is difficult to use SVMs to do feature selection

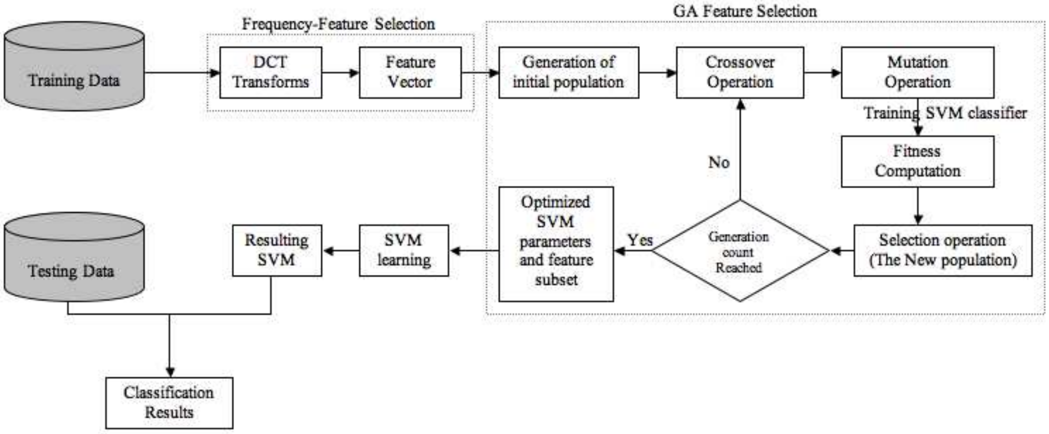

Figure 2: The general process for feature subset selection and classification.

directly. Some researchers have proposed approximations to SVM for feature selection by first training the SVM using the whole training set, and then computing approximations to reduce the number of features.

4 Overview of the Proposed Method

The main steps of the proposed method are as follows:

1. FFS extraction using DCT [4].
2. Using Genetic Algorithms, in order to generate both the optimal feature subset and SVM parameters at the same time [5].
3. Classification of novel images.

Fig. 2 presents the general schema of feature selection and classification process. Firstly, a population of possible frequency features subset is genetically evolved, these features seems to be most useful to a particular classification problem from all those available, it can be explain that they contain only highly informative and non-redundant features, which significantly improve classification. The genetic evolution is guided using the proposed fitness criterion, the quality of a given chromosome is proportional to the information gain measure computed using the dataset records retrieved from the training dataset, the chromosome comprises three parts, C, kernel parameter, and the features mask. The result is finally validated using a new test dataset. In fact, the basic idea here consists in using a GA to discover "good" subsets of genes, the goodness of a subset being evaluated by SVM classifier.

Using these methods we obtained three benefits, the first one that computational complexity is reduced as there is smaller number of inputs. Often, a secondary benefit found is that the accuracy of the classifier increases, and the last one is to remove the extra features (i.e like noise, obscuring other features from the learning algorithm) from a feature set, like unnecessary information showed in Fig. 4.

5 Experiment results and comparison

5.1 The Dataset

To assess the robustness of our method against different facial expressions, lighting conditions and pose, we have collect grey-scale face images from two different face database available in the public domain, ORL face database [1] and Yale face database [2].

Figure 3: Some samples from the used face database (ORL+Yale).

Face images selected are near frontal and contain

[1] http://www.cl.cam.ac.uk/Research/DTG/attarchive:pub/data /att_faces.zip

[2] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

variations in pose, illumination and expression. Eyebrows, eyes, nose, lips and surrounding area of face image contribute maximum in face recognition. So scale normalization of face images of data sets is carried out using the cropping phenomena which eliminate the unnecessary information from image and retain only internal structures. All the faces are then scaled to the size 48×48 pixels, aligned according to the eye positions. Sample images from the face databases are shown in Fig. 3. There are 330 subjects with 10 images per subject for a total of $3,300$ images. The entire face database (ORL + Yale) is divided into two parts, six images of each subject are used to construct the training data and the remaining images are used for testing.

5.2 DCT based Feature Extraction

We have performed a number of experiments [4] in order to demonstrate the performance of the DCT based Feature Extraction on gray-scale images. In our study, DCT is used to extract pertinent information which represent low frequency in each block. The local information of a candidate face can be obtained by using block-based DCT as follows: a face image is divided into blocks of 8 by 8 pixels size without overlapping. Each block is then represented by its DCT coefficients. From the obtained DCT coefficients only a small, *generic* feature set is retained in each block (see Figure 4). Ekenel et al. [19] has proved that the highest information necessary to achieve high classification accuracy is contained in the first low frequency DCT coefficients via zigzag scanning.

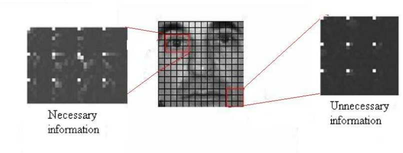

Figure 4: Illustration of the effects of the blockbased DCT for local appearance based face representation.

5.3 GA and Classifier parameters

We used the GA approach to select a set of good FFS for SVMs classifier, the polynomial kernel has been found in our simulations to outperform linear and RBF kernel functions. In the present work, the library LIBSVM [3] was used with a 10-fold cross-validation on the training data.

In the GA using parameters dressed in table 1, pairs of (C,d) are tried and the one with the best cross-validation accuracy is chosen. In the classification step, we use SVM with Polynomial kernel functions with the best parameters which are obtained by simulation, while varying the dimensionality of the generation.

Parameter	**Default value**	**Signification**
Population size	30	Number of chromosomes created in each generation
Crossover rate	0.8	Probability of crossover
Mutation rate	0.1	Probability of mutation
Number of generations	20	Maximum number of generations

Table 1: Parameters set used for the genetic process

5.4 Evaluation and comparison of Feature Selection Techniques

In this section, we perform comprehensive experiments on face image database (ORL + Yale) to compare the GA-SVM selection algorithm with the different state of the art approaches discussed above: Ranking algorithm (Rank), Minimum Redundancy Maximum Relevance criterion(mRMR-D and mRMR-Q), Conditional Mutual Information Maximization criterion (CMIM) and Interaction Gain criterion (IGFS).

The accuracy of classification (recognition rate) relatively to the step by step introduction of the variables is computed and the evolution of the recognition rate using different feature selection algorithm is reported in Figure 5.

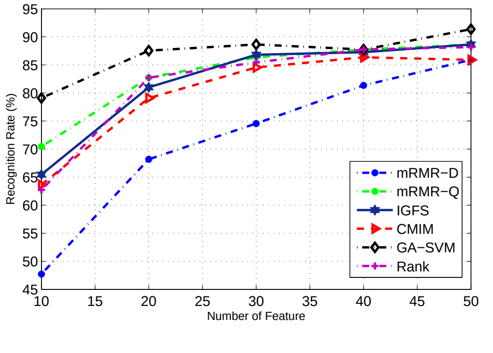

Figure 5: A comparison of feature selection methods in the context of SVM

As show in Figure 5, the GA-SVM can give the best

[3]http://www.csie.ntu.edu.tw/ cjlin/libsvm

result of 79.1% accuracy using only 10 frequency feature. We obtain 87.73% classification performance in GA-SVM, IGFS and mRMR-Q with 40 frequency feature. Moreover, GA-SVM is better than the other feature selection algorithms on the different length of feature subsets. This obtained result proves the strength of the GA-SVM compared with the other feature selection algorithms based on Mutual Information. In addition, GA-SVM gives 91.36% with only 50 frequency features compared to using the whole information without GA selection which gives $89,1\%$ with the same SVM parameters.

The analysis of this graph allowed us to take out the following results:

1. The measures based on the mutual information can be used for performing feature selection for the problem of face recognition, specially, mRMR with Quotient (mRMR-Q) to improve a best mutual information technique compared to others;

2. The performance of mRMR-D is less than the others ;

3. GA-SVM achieve higher recognition rate using only few frequency feature subset;

6 Conclusions

In our work, we proposed a framework for face recognition based feature extraction using DCT and GA-SVM for frequency feature selection. GA-SVM with parameter optimisation is proved to be effective in selecting FFS and significant fitness even if the sample set is very small. To assess the effectivness of our proposed method, we performed a critical comparison with several mutual information methods. Our proposed method outperforms all of the other methods based on mutual information criteria.

References

[1] Ahmad, A. and Dey, L. A feature selection technique for classificatory analysis. *Pattern Recognition Letter*, 26(1):43–56, 2005.

[2] Ahmed, N., Natarajan, T., and Rao, K. Discrete cosine transform. *IEEE Transactions Computers*, 23:90–94, 1974.

[3] Akadi, A. E., Ouardighi, A. E., and Aboutajdine, D. International journal of computer science and network security. *In Proceedings of the 13th international conference on machine learning, San Francisco, CA*, 8(4):116–121, 2007.

[4] Amine, A., Ghouzali, S., Rziza, M., and Aboutajdine, D. Investigation of feature dimension reduction based dct/svm for face recognition. *IEEE Symposium on Computers and Communications (ISCC'08)*, 2008.

[5] Amine, A., Rziza, M., and Aboutajdine, D. Svm-based face recognition using genetic search for frequency-feature subset selection. *ICISP, LNCS 5099*, pages 321–328, 2008.

[6] Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Transaction Neural Networks*, 5(4):537–550, 1994.

[7] Bishop, C. Neural networks for pattern recognition. *Oxford University Press*, 1995.

[8] Bradley, P. and Mangasarian, O. Feature selection via concave minimization and support vector machines. *In Proceedings of the 13th international conference on machine learning, San Francisco, CA*, pages 82–90, 1998.

[9] Brill, F., Brown, D., and Martin, W. Fast genetic selection of features for neural network classifers. *IEEE Transaction on Neural Networks*, 3(2):324–328, 1992.

[10] Burges, J. C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[11] Caruana, R. and Freitag, D. Greedy attribute selection. *Proc. of the 11th Internat. Conf. on Machine Learn. New Brunswick, NJ, USA*, pages 28–36, 1994.

[12] Chellappa, R., Wilson, C. L., and Sirohey, S. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, 1995.

[13] Chen, C. Statistical pattern recognition. *Spartan Books, Washington, DC*, pages 82–90, 1973.

[14] Congdon, C. A comparison of genetic algorithm and other machine learning systems on a complex classification task from common disease research. *PhD thesis, University of Michigan, USA*, 1995.

[15] Cover, T. and Thomas, J. Elements of information theory. *Wiley, New York*, 1991.

[16] Dash, M. and Liu, H. Feature selection for classification. *Intell. Data Analysis*, 1(3):131–156, 1997.

[17] Doak, J. An evaluation of feature selection methods and their application to computer security. *Technical Report CSE-92-18 , University of California at Davis.*, 1992.

[18] Duch, W., Winiarski, T., Biesiada, J., and Kachel, A. Feature selection and ranking filters. *International Conference on Artificial Neural Networks (ICANN) and International Conference on Neural Information Processing (ICONIP)*, pages 251–254, 2003.

[19] Ekenel, H. and Stiefelhagen, R. Local appearance based face recognition using discrete cosine transform. *EUSIPCO 2005, Antalya, Turkey*, 23(7), 2005.

[20] Faraoun, K. M. and Rabhi, A. Data dimensionality reduction based on genetic selection of feature subsets. *Journal of Computer science*, pages 36–46, 2007.

[21] Fleuret, F. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, 2004.

[22] Friedman, J. Another approach to polychotomous classification. *Technical report, Department of Statistics, Stanford University*, 1996.

[23] Goldberg, D. Genetic algorithms in search, optimization, and machine learning. *Addison Wesley*, 1989.

[24] Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *JMLR*, 3:1157–1182, 2003.

[25] Hafed, Z. M. and Levine, M. D. Face recognition using the discrete cosine transform. *International Journal of Computer Vision*, 43(3):167–188, 2001.

[26] Holland, J. Adaptation in nature and artificial systems. *MIT Press*, 1992.

[27] Jain, A. and Zongker, D. Feature selection: Evaluation, application, and small sample performance. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(2):153–158, February 1997.

[28] Jakulin, A. Attribute interactions in machine learning. *Master's thesis, University of Ljubljana, Faculty of Computer and Information Science*, 2003.

[29] Kohavi, R. and John, G. Wrappers for feature subset selection. *Artificial Intell.*, pages 273–324, 1997.

[30] Koller, D. and Sahami, M. Towards optimal feature selection. *ICML-96, Bari, Italy*, pages 87–95, 1996.

[31] Peng, H. and Long, F. An efficient max-dependency algorithm for gene selection. *In: 36th Symposium on the Interface: Computational Biology and Bioinformatics*, 2004.

[32] Raymer, M. L., Punch, W. F., Goodman, E. D., Kuhn, L. A., and Jain, A. Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation*, 4(2):164–171, 2000.

[33] Shannon, C. and Weaver, W. The mathematical theory of communication. *University of Illinois Press, Urbana, IL*, 1949.

[34] Shneider, M. and Abdel-Mottaleb, M. Exploiting the jpeg compression scheme for image retrieval. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18(8), 1996.

[35] Somol, P., Pudil, P., and Kittler, J. Fast branch and bound algorithms for optimal feature selection. *IEEE Trans. Pattern Anal. Machine Intell*, 26(7):900–912, 2004.

[36] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. Feature selection for support vector machines. *Advances in Neural Information Processing Systems*, 2000.

[37] Yang, J. and Honavar, V. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems*, 13(2):44–49, 1998.

[38] Yonghua, X., Lokesh, S., and Hans, B. Block dct vectors construction for face retrieval based on genetic algorithm. *The 3rd International Conference on Natural Computation (ICNC'07)*, 3, 2007.

[39] Yu, L. and Liu, H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.