Journal of Information Hiding and Multimedia Signal Processing ©2014 ISSN 2073-4212
Ubiquitous International Volume **5**, Number **4**, October **2014**

Face Feature Selection with Binary Particle Swarm Optimization and Support Vector Machine

Hongtao Yin, JiaQing Qiao, Ping Fu, XinYuan Xia

Department of Automatic Test and Control
Harbin Institute of Technology, Harbin, China.
yinht@hit.edu.cn

Received April, 2013; revised April, 2014

Abstract. *A face feature selection and recognition method based on BPSO and SVM-Wrapper model is presented. To solve the problem that DCT coefficients dimension is higher for face recognition, we design a SVM-Wrapper model based on BPSO. In the process of training SVM, the cross-validation is used to training samples, and the recognition accuracy is used for defining the fitness function of BPSO feature selection algorithm. The fitness function is used to guide the BPSO algorithm to search the optimal feature subset. The experiments on ORL databases show that the improved method is effective.*
Keywords: Face recognition, Support vector machine, Feature selection, Discrete cosine transform, Binary particle swarm optimization.

1. **Introduction.** Face recognition is one of the most interesting and challenging areas in computer vision and pattern recognition. Although research in the field of face recognition is active over 30 years and considerable successes in face recognition systems have been achieved, there are still some unsolved problems. Illumination variation, rotation and facial expression are the basic existing challenges in this area.

Many approaches have been developed by researchers for face recognition problem. An excellent face recognition method should consider what features are used to represent a face image and how to classify a new face image based on this representation. Current feature extraction methods can be classified into signal processing and statistical learning methods. On signal processing based methods, feature extraction based Gabor wavelets are widely used to represent the face image [1][2]. On the statistical learning based methods, the dimension reduction methods are widely used in the past works [3], as the famous face recognition method, Principal Component Analysis (PCA) has been widely studied, and the PCA and liner discriminant analysis (LDA)[4][5] are widely used among the dimensionality reduction methods [6]. Recently kernel based nonlinear feature extraction methods are applied to face recognition [7], such as kernel principal component analysis (KPCA)[8], kernel discriminant analysis (KDA)[9]. LDA is to find the optimal projection matrix with Fisher criterion through considering the class labels. Recently, researchers proposed some other manifold algorithms such as Locally Linear Embedding (LLE) [10] and Locality Preserving Projection (LPP) [11]. Many improved LPP algorithms were proposed in recent years [12-14].

Discrete cosine transform (DCT) is a signal processing method which transform the data from time domain to frequency domain. Discrete Cosine Transform is a real number domain transformation, and the transform coefficient distribution is more concentrated. The main information of an image is the low-frequency information. The main information

of the image is concentrated in the low frequency after the discrete cosine transform. So DCT is widely used in the voice and image data compression field. Also it can be extracted the DCT coefficients feature for classification.

DCT has been employed in face recognition for dimensionality reduction [15]. The advantage of DCT is that it is data independent and it can be implemented using a fast algorithm. Nevertheless, only limited low-frequency coefficients are used as features if the discrete cosine transform is employed for direct dimensionality reduction [16]. D. Ramasubramanian et al.[17] proposes a face recognition method using DCT combined with LDA. DCT does not compress the data itself, and the source data of the image is only mapped to another domain. How to select the most effective DCT coefficients as the identification feature in the new data field becomes a key issue. Saeed Dabbaghchian et al.[18] and Yin et al.[19] propose using separability measure to select the DCT coefficients. It searches the coefficients which have more power to discriminate different classes are better than the others. They are able to find the DCT coefficients on each dimension database. Both methods can select the individual DCT coefficients which have the best discriminant ability. However, the discriminant ability of the feature vector combined with the selected DCT coefficients is not strong necessarily. In this paper, to solve the problem that DCT coefficients dimension is higher for face recognition, we propose a novel method for face feature selection and recognition that is based on binary particle swarm optimization algorithm and support vector machine. Firstly, we apply discrete cosine transform to extract the DCT coefficients of face images. Secondly, we apply support vector machine and particle swarm algorithm to construct a Wrapper feature selection model, and then this feature selection model is used to select recognition features.

The rest of this paper is organized as follows. Section2 describes some related works. Our proposed BPSO SVM-Wrapper model is discussed elaborately in Section 3. Section 4 presents the results of our experimental studies including the experimental methodology, experimental results, and the comparison with other existing algorithms. Finally, Section5 concludes the paper with a brief summary and a few remarks.

2. **Related work.**

2.1. **Discrete cosine transform of an image.** The DCT has been widely applied to solve numerous problems among the digital signal processing community. In particular, many data compression techniques employ the DCT, which has been found to be asymptotically equivalent to the optimal Karhunen-Loeve Transform (KLT) for signal decorrelation. Mathematically, the DCT is a one-to-one mapping between the spatial and spectral domains. An image is transformed to its spectral representation by projection into a set of orthogonal 2-D basis functions. The amplitudes of these projections are called the DCT coefficients, which are the output of the transform. Given an input $M \times N$ image $f(x, y)$, its DCT, $C(u, v)$ is obtained by the following equation

$$C(u,v) = a(u)\, a(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cos \frac{(2x-1)\, u\pi}{2M} \cos \frac{(2y-1)\, v\pi}{2N} \tag{1}$$

Where$u = 0, 1, \cdots, M-1$ $v = 0, 1, \cdots, N-1$ and $a(u)$ $a(v)$ are defined by:

$$a(u) = \begin{cases} \sqrt{1/M} & u = 0 \\ \sqrt{2/M} & u = 1, 2, \cdots, M-1 \end{cases} \tag{2}$$

$$a(v) = \begin{cases} \sqrt{1/N} & v = 0 \\ \sqrt{2/N} & v = 1, 2, \cdots, N-1 \end{cases} \tag{3}$$

Where, x and y are spatial coordinates in the sample domain while u, v are coordinates in the transform domain.

For an $M \times N$ face image, we have an $M \times N$ DCT coefficient matrix covering all the spatial frequency components of the image. Figure 1 shows an $M \times N$ face image and its DCT coefficients.

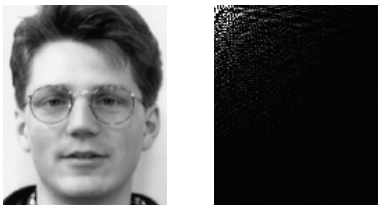

FIGURE 1. Face image and its DCT coefficients

The DCT is applied to the entire image to obtain the frequency coefficient matrix of the same dimension. Therefore coefficients selection, the second stage of the feature extraction, is an important part of the feature extraction process and strongly influences the recognition accuracy.

The conventional DCT coefficient selection approaches select the fixed elements of the DCT coefficients matrix. Most of the conventional approaches select coefficients in a zigzag manner or by zonal masking. The DCT coefficients with large magnitude are mainly located in the upper-left corner of the DCT coefficients matrix. It can be observed that a large amount of information about the original image is stored in the upper-left corner of the DCT coefficients matrix. They are the low spatial frequency DCT components in the image.

These are good criterions in case of compression is not recognition. Since the discrimination power of all the coefficients is not the same and some of them are discriminant than the others. Yin et al. [18, 19] proposed using separability criterion to select DCT coefficients. It searches for the coefficients which have more power to discriminate different classes are better than the others. In our method, the separability criterion method is used for pre-selection of the DCT coefficients.

2.2. **Binary Particle Swarm Optimization.** Particle swarm optimization (PSO) is a population-based evolutionary computation technique developed by Kennedy and Eberhart in 1995. PSO simulates the social behavior of organisms, i.e., birds in a flock or fish in a school. This behavior can be described by a swarm intelligence system. In PSO, each solution can be considered as an individual particle in a given search space, which has its own position and velocity. During movement, each particle adjusts its position by changing its velocity based on its own experience, as well as the experience of its companions, until an optimum position is reached by itself and its companions [21]. All of the particles have fitness values based on the calculation of a fitness function. Particles are updated by following two parameters called *pbest* and *gbest* at each iteration. Each particle is associated with the best solution (fitness) particle has achieved so far in the search space. This fitness value is stored, and represents the position called *pbest*. The value *gbest* is a global optimum value for the entire population.

Many optimization problems occur in a space featuring discrete, qualitative distinctions between variables and levels of variables. To extend the real-value version of PSO to a binary space, Kennedy and Eberhart proposed a binary version of the PSO method (BPSO). In a binary search space, a particle may move to near corners of a hypercube

by flipping various number of bits; thus, the overall particle velocity may be described by the number of bits changed per iteration [22].

The position of each particle is represented in binary string form by $X_i = \{x_{i1}, x_{i2}, ..., x_{iD}\}$ which is randomly generated. The bit values 0 and 1 represent a non-selected and selected feature, respectively. The velocity of each particle is represented by $V_i = \{v_{i1}, v_{i2}, ..., v_{iD}\}$ (i is the number of particles, and D is the number of dimensions (features) of a given data set). The initial velocities in particles are probabilities limited to a range of [0, 1]. The best solution ever encountered by individual agents during the search process is stored as local best solution (*pbest*) for that particular agent and the solution with the highest fitness of all the solutions in the swarm is stored as the global best solution (*gbest*): The rest of the agents of the swarm update their velocity v and position x using the information associated with *pbest* and *gbest* as shown in equations (4) and (5), respectively.

$$v_{id}^{n+1} = v_{id}^{n} + c_1 r_1 (p_{id}^{n} + x_{id}^{n}) + c_2 r_2 (p_{gd}^{n} + x_{id}^{n}) \tag{4}$$

$$x_{id}^{n+1} = x_{id}^{n} + v_{id}^{n+1} \tag{5}$$

Where, the constants c_1 and c_2 are positive acceleration constants, whereas r_1 and r_2 are uniformly distributed random numbers in [0, 1]. A velocity-range [-$v_{\max}$, $v_{\max}$] is defined to ensure that the agents do not fly out of the swarm.v_{id}^{n+1} and x_{id}^{n+1} represent the updated velocity and position, respectively, for ith agent in the swarm at *(n+1)*th iteration. p_{id}^{n} and p_{gd}^{n} represent the *pbest* for ith agent and *gbest* of the entire swarm at nth iteration, respectively.

The velocity update v_{id}^{n+1} with the above formulation can be calculated using equation (4) followed by updating x_{id}^{n+1} as equation (5). It is still a continuous valued x_{id}^{n+1} which needs to be in binary form. This is addressed by calculating an intermediate variable $s(v_{id}^{n+1})$ from continuous valued v_{id}^{n+1} through sigmoid limiting function as shown in (6).

$$s(v_{id}^{n+1}) = \frac{1}{1 + e^{-v_{id}^{n}}} \tag{6}$$

The values of $s(v_{id}^{n+1})$ which is monotonically increasing function, can be interpreted as the probability of changing the current state of the ith agent at the nth iteration. Higher values of $s(v_{id}^{n+1})$ indicate higher possibility of changing a bit to 1 and vice-versa. At the same time it squashes the velocity range [-$v_{\max}$, $v_{\max}$] within the range [0, 1] and eliminates the requirement of boundary handling techniques as required in case of continuous domain PSO. The bit values of the solutions are updated probabilistically with $s(v_{id}^{n+1})$ using equation (7).

$$f\ (rand() < s(v_{id}^{n}))\quad then\ x_{id}^{n+1} = 1;\quad else\quad x_{id}^{n+1} = 0 \tag{7}$$

2.3. **Support Vector Machine.** Support Vector Machines (SVM) [23] is an effective classification method with significant advantages such as the absence of local minima, an adequate generalization to new objects, and a representation that depends on few parameters. A Support Vector Machine performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. Using a kernel function, SVM is an alternative training method for polynomial, radial basis function and multi-layer perception classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training. The goal of SVM model is to find the optimal hyperplane that separates the clusters of vectors into two sides of the plane, where the same category cases of the target variable on the same side.

It works well by using a hyperplane to separate the feature vectors into two groups when there are only two target categories. For the multi-class classifier, there are two major multi-class SVM classification strategies: one-against-all and one-against-one [24]. For the one-against-all strategy, multiple binary SVM decision functions are constructed. Every decision function is trained by labeling all of the examples which are in this class with positive labels, and all of the examples not in this class with negative labels. A new sample is classified into the class which has the largest decision function. For one-against-one strategy multiple classifiers are constructed, and each classifier is trained with two different classes. A new sample is classified into the majority class voted by all of the indicator functions.

3. **BPSO SVM-Wrapper model.** Feature selection is of considerable importance in classification. Feature selection addresses the dimensionality reduction problem by determining a subset of available features to build a good model for classification or prediction, which is a combinatorial problem in the number of original features.

Support Vector Machines is an effective classification method with significant advantages such as the absence of local minima, an adequate generalization to new objects, and a representation that depends on few parameters. This method, however, does not directly determine the importance of the features used.

In this section, a new feature selection method based on BPSO and SVM are considered and a new efficient approach is proposed. The architecture for the BPSO-based SVM classifier is implemented in this paper. A more detail information flows of the PSO-SVM feature selection model is illustrated in Figure 2.

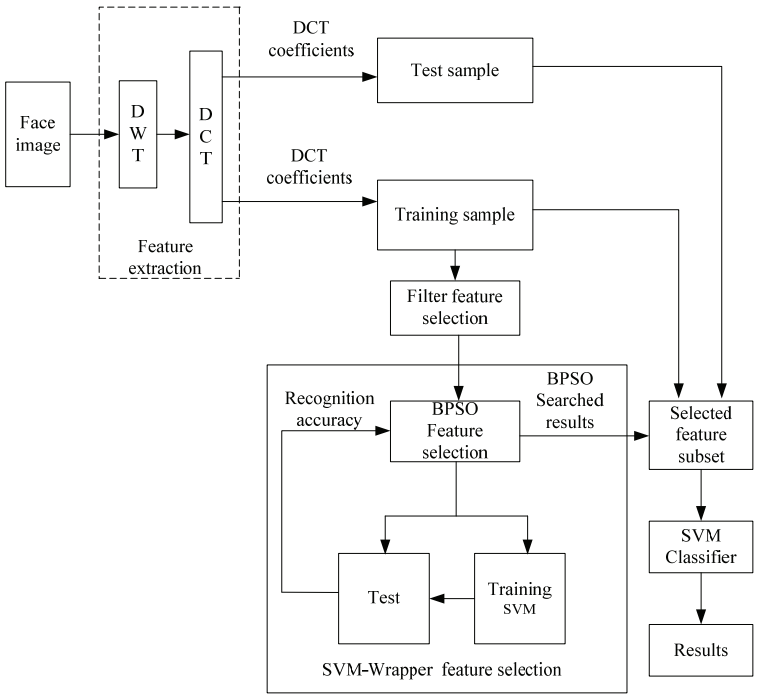

FIGURE 2. The architecture of our method

DCT feature extraction consists of two stages. Dimension of the DCT coefficient matrix is the same as the input image. In fact the DCT, by itself, does not decreased at a dimension. In the first stage, it compresses most signal information to the low frequency

by wavelet transform. The DCT is applied to the low frequency sub-band image to obtain the DCT coefficients, and then some of the coefficients are selected to construct feature vectors in the second stage.

In the initial experimentation, we observed that it can not get a good recognition performance if the classification features is obtained by searching all the DCT coefficients using discrete PSO algorithm. In general, the DCT coefficients are divided into three bands, namely low frequencies, middle frequencies and high frequencies. Low frequencies and middle frequencies coefficients contain useful information and construct the basic structure of the image. High frequencies represent noise and small variations. From the above discussion, it seems that the low frequencies and middle frequencies coefficients are more suitable candidates in face recognition. Therefore the search space of BPSO algorithm is selected in the low frequency and middle frequencies coefficients. This is the filter feature selection shown in the figure 2.

In Wrapper feature selection model, the search strategy of feature subset is implemented by BPSO algorithmand the SVM classifier. In the search process, the cross-validation method is used to test the classification ability of features. Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one is used to learn or train a model and the other is used to validate the model. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. The basic form of cross-validation is k-fold cross-validation. Other forms of cross-validation are special cases of k-fold cross-validation or involve repeated rounds of k-fold cross-validation.

In our experiments leave-one-out cross-validation method is used. Leave-one-out cross-validation is a special case of k-fold cross-validation where k equals the number of instances in the data. In other words, all the data except for a single feature are used for training and the model is tested on that single feature. After a cycle using this training set can obtain the average recognition rate of the SVM classifier, which is used to form the fitness function. The fitness function is shown in the following formula,

$$Fitness = P \times Accuracy - Dimensions \tag{8}$$

Where, *Accuracy* is the average recognition rate, P is the weight of *Accuracy*, *Dimensions* is the characteristic dimension of the final choice. The purpose of feature selection is to improve the recognition rate and reduce the number of feature dimensions. But sometimes they are not unified. If the recognition rate is more important, P can take a larger value.

The main purpose of the fitness function is to select the features that can improve the recognition rate. Under the precondition of ensuring the classification accuracy, the redundant feature in the feature subset are removed, i.e. excluding those that do not affect the test recognition rate, thereby reducing the dimension of the selected features. Through the SVM-Wrapper model, we can find an optimal combination of some features. And then the recognition features are selected according to this combination.

4. **Experimental Results.** In order to test the performance of the proposed method, some experiments are performed on two face database. One is ORL face database which contains a set of face images taken at the Olivetti Research Laboratory (ORL) in Cambridge University, U.K. There are 400 images of 40 individuals in it. For some subjects, the images were taken at different times, which contain quite a high degree of variability in lighting, facial expression, pose, and facial details. Another is Bern face database. This database involves 300 frontal facial images, with 10 images of 30 individuals. The size of

each image is 320×214 with 256 gray levels. Each image is scaled down to the size of 92×112 pixels.

In the experiments, we randomly select five images from each subject to construct the training data set, the remaining images are used as the test images. Each experiment is repeated 20 times. The BPSO algorithm uses 40 particles, 25 iterations. In order to highlight the importance of the recognition rate, in the formula of the fitness function, the weight P of *Accuracy* equals 100000. The value of inertia weight w equals 0.7 and the value of acceleration constants c_1 and c_2 both equal 1.5. The SVM uses the LIBSVM MATLAB toolbox. The polynomial is selected as the kernel function of SVM.

The average recognition rates of the different training set dimension are, respectively, summarized in Table 1 and Table 2. In where, Original DCT means the DCT coefficients of face image as face feature vector without the feature selection. DWT + DCT means the DCT coefficients of the face image after the second-order wavelet transform as a face feature vector without the feature selection. Filter1 means the selected DCT coefficient matrix of the low-frequency portion of the upper left corner as the feature vectors of 120-dimensional (10 rows 12 columns). Filter2 means the selected DCT coefficient matrix of the low-frequency portion of the upper left corner as the feature vectors of 80-dimensional (8 rows 10 columns). Filter3 means selected DCT coefficient matrix of the low-frequency portion of the upper left corner as the feature vectors of 42-dimensional (6 rows 7 columns). SVM-Wrapper1, SVM-Wrapper2 and SVM-Wrapper3 mean based on the SVM-wrapper face feature selection and identification method for the above Filter method, respectively.

Table 1. Simulation results on ORL face database

Algorithms	Recognition accuracy (%)	Feature dimensions
Original DCT	94.80	10304
DWT+DCT	95.03	644
Filter1	95.43	120
SVM-Wrapper1	96.03	58
Filter2	95.45	80
SVM-Wrapper2	97.33	41
Filter3	95.98	42
SVM-Wrapper3	97.60	31

Table 2. Simulation results on Bern face database

Algorithms	Recognition accuracy (%)	Feature dimensions
Original DCT	93.96	10304
DWT+DCT	94.72	644
Filter1	94.89	120
SVM-Wrapper1	95.45	61
Filter2	95.12	80
SVM-Wrapper2	95.86	43
Filter3	95.36	42
SVM-Wrapper3	96.79	33

In Table 1 and Table 2, we can find that the DCT method obtains a recognition rate of 94.80% on ORL database and a recognition rate of 93.96% on Bern database although it use all features. The DWT+DCT method and the Filter obtain more recognition results using a less features than DCT method. However the SVM-Wrapper method obtain the best recognition accuracy using a less features than Filter method.

For the purpose of comparison, we compare our method with several face recognition methods based on the discrete cosine transform. In the literature [15], the low-frequency coefficients of DCT coefficient matrix in the upper left corner within a square is directly used for classification. In the literature [20], the face image is processed by discrete cosine transform, and then selecting the low frequency coefficients of the DCT coefficient matrix in the upper left corner within a square to extract the face features by linear discriminant analysis. In the literature [19], the separability are calculated each dimensional feature according to separability criterion, and then the features with better separability are used for classification.

Table 3. Performance comparison on ORL face database

	Recognition accuracy	Feature dimensions
DCT[15]	94.97%	64
DCT+LDA[20]	96.65%	49
Separability Criterion[19]	97.00%	25
SVM-wrapper	97.60%	31

As can be seen from Table 3, the proposed method use only 31 DCT coefficients obtain the best recognition results. We also observe from experimental results the number of DCT coefficients using classification is not proportional to the obtaining recognition rate. Using a large number of DCT coefficients can not obtain the best recognition rate. In the reconstruction of a face image, it is necessary for accurately reconstructing the original image to use as much as possible DCT coefficients, but recognition does not require much DCT coefficients. The disadvantage of our approach compared to other methods is high computational complexity in the training stage.

5. **Conclusions.** In this paper, for the DCT coefficients selection, a face feature selection and face recognition method based on support vector machine and particle swarm is proposed. From the view point of select effective features, we have established a Wrapper feature selection model in which support vector machine and particle swarm search algorithm are the core. The Cross-Validation method is used to define the fitness function of BPSO feature selection algorithm. Simulation results demonstrate the effectiveness of the method.

REFERENCES

[1] C. Liu and H. Wechsler, Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition *IEEE Trans. Image Processing*, vol. 11, no. 4, pp. 467-476, 2002.

[2] H. Zhang, B. Zhang, W. Huang and Q. Tian, Gabor Wavelet Associative Memory for Face Recognition, *IEEE Trans. Neural Networks*, vol. 16, no. 1, pp. 275-278, 2005.

[3] Y. Xie, L. Setia and H. Burkhardt, Face Image Retrieval Based on Concentric Circular Fourier-Zernike Descriptors, *International Journal of Innovative Computing, Information and Control*, vol. 4, no. 6, pp. 1433-1444, 2008.

[4] J. B. Li, J. S. Pan and S. C. Chu, Kernel Class-wise Locality Preserving Projectio, *Information Sciences*, vol. 178, no. 7, pp. 1825-1835, 2008.
[5] J. S. Pan, J. B. Li and Z. M. Lu, Adaptive Quasiconformal Kernel Discriminant Analysis, *Neurocomputing*, vol. 71, pp. 2754-2760, 2008.
[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.
[7] A. Ruiz and P. E. Lopez-de-Teruel, Nonlinear kernel-based statistical pattern analysis, *IEEE Trans. Neural Networks*, vol.12, no.1 , pp. 16-32, 2001.
[8] H. Sahbi, Kernel PCA for similarity invariant shape recognition, *Neurocomputing*, vol.70 , no.9 , pp.3034-3045, 2007.
[9] Q. Liu, H. Lu, and S. Ma, Improving Kernel Fisher Discriminant Analysis for Face Recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no.1, pp. 42-49, 2004.
[10] S. T. Roweis and L. K. Saul, Nonlinear dimensionality deduction by locally linear embedding, *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000.
[11] X. He and P. Niyogi, Locality preserving projections, *Proc. Conf. Advances in Neural Information Processing Systems*, pp.585-591, 2003.
[12] Z. Zheng, F. Yang, W. Tan, J. Jia, and J. Yang, Gabor feature-based face recognition using supervised locality preserving projection, *Signal Processing*, vol. 87, no. 10, pp.2473-2483, 2007.
[13] L. Zhu and S. Zhu, Face recognition based on orthogonal discriminant locality preserving projections, *Neurocomputing*, vol. 70, no. 7, pp. 1543-1546, 2007.
[14] D. Cai, X. He, J. Han, and H. J. Zhang, Orthogonal Laplacianfaces for face recognition, *IEEE Transaction on Image Processing*, vol. 15, no. 11, pp. 3608-3614, 2006.
[15] Z. M. Hafed, M. D. Levine, Face Recognition Using the Discrete Cosine Transform, *International Journal of Computer Vision*, vol.43, no.3, pp. 167-188, 2001.
[16] Weilong Chen, Meng Joo Er, Shiqian Wu, PCA and LDA in DCT domain, *Pattern Recognition Letters*, vol.26, no.7, pp. 2474-2482, 2005.
[17] D. Ramasubramanian, Y.V. Venkatesh, Encoding and Recognition of Faces Based on the Human Visual Model and DCT, *Pattern Recognition*, vol. 34, no.12, pp. 2447-2458, 2001.
[18] Saeed Dabbaghchian, Masoumeh P.Ghaemmaghami, Ali Aghagolzadeh, Feature extraction using discrete cosine transform and discrimination power analysis with a face recognition technology, *Pattern Recognition*, vol.43, no.11, pp. 1431-1440, 2010.
[19] Yin Hongtao, Fu Ping, Sha Xuejun, Face recognition based on DCT and LDA, *Tien Tzu Hsueh Pao/Acta Electronica Sinica*, vol. 37, no. 10, pp:2211-2214, 2009.
[20] Zhang, Yankun; Liu, Chongqing, A Novel Face Recognition Method Based On Linear Discriminant Analysis, *Journal of infrared and millimeter waves*, vol. 22 , no. 5, pp. 327-330, 2003.
[21] R. Mendes, J. Kennedy, J. Neves, The informed particle swarm: simpler, maybe better, *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 204-210, 2004.
[22] J. Kennedy, R.C. Eberhart, A discrete binary version of the particle swarm algorithm, *Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, Piscataway, NJ*, pp. 4104-4109, 1997.
[23] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, NY, USA, 1995.
[24] U. KreBel, Pairwise Classification and Support Vector Machines, Advances in Kernel Methods Support Vector Learning, *MIT Press*, Cambridge, MA, USA, pp. 254-268, 1999.