Comparative Study of PCA, ICA, LDA using SVM Classifier

Anissa Bouzalmat
Sidi Mohamed Ben Abdellah University
Department of Computer Science
Faculty of Science and Technology
Route d'Imouzzer B.P.2202 Fez 30000 Morocco
anissabouzalmat@yahoo.fr

Jamal Kharroubi
Sidi Mohamed Ben Abdellah University
Department of Computer Science
Faculty of Science and Technology
Route d'Imouzzer B.P.2202 Fez 30000 Morocco
jamal.kharroubi@yahoo.fr

Arsalane Zarghili
Sidi Mohamed Ben Abdellah University
Department of Computer Science
Faculty of Science and Technology
Route d'Imouzzer B.P.2202 Fez 30000 Morocco
a.zarghili@ieee.ma

***Abstract*—Feature representation and classification are two key steps for face recognition. We compared three automated methods for face recognition using different method for feature extraction: PCA (Principle Component Analysis), LDA (Linear Discriminate Analysis), ICA (Independent Component Analysis) and SVM (Support Vector Machine) were used for classification. The experiments were implemented on two face databases, The ATT Face Database [1] and the Indian Face Database (IFD) [2] with the combination of methods (PCA+ SVM), (ICA+SVM) and (LDA+SVM) showed that (LDA+SVM) method had a higher recognition rate than the other two methods for face recognition.**

***Index Terms*—Face Recognition, SVM, LDA, PCA, ICA.**

I. Introduction

Face Recognition is a term that includes several sub-stages as a two step process: Feature extraction and classification.

Feature extraction for face representation is one of central issues to face recognition systems, it can be defined as the procedure of extracting relevant information from a face image.

There are many feature extraction algorithms, most of them are used in other areas than face recognition.

Researchers in face recognition have used, modified and adapted many algorithms and methods to their purpose . For example, Principle component analysis (PCA) was applied to face representation and recognition [3, 4, 5].

The PCA method [5] is obviously of advantage to feature extraction, but it is more suitable for image reconstruction because of no consideration for the separability of various classes. Aiming at optimal separability of feature subspace, LDA (Linear Discriminate Analysis) can just make up for the deficiency of PCA [6]. ICA (Independent Component Analysis) is a method that finds better basis by recognizing the high-order relationships between the pixels images [7], once the features are extracted, the next step is to classify the image .A large margin classifiers are proposed recently in machine learning such as Support Vector Machine (SVM) [8]. The method was used in this step is SVM (Support Vector Machines) which have been developed in the frame work of statistical learning theory, and have been successfully applied to a number of applications, ranging from time series prediction, to face recognition, to biological data processing for medical diagnosis [9,10]. VC (Vapnik-Chervonenkis) dimension theory and SRM (Structural Risk Minimization) principle based SVM can well resolve some practical problems such as small sample size, nonlinear, high dimensional problems, etc. [11,12] .

In this paper SVMs were used for classification using different method for feature extraction: PCA, LDA, ICA, the experiments were implemented on two face databases, The ATT Face Database [1] and the Indian Face Database (IFD) [2] .

The face recognition system is shown as Fig. 1.

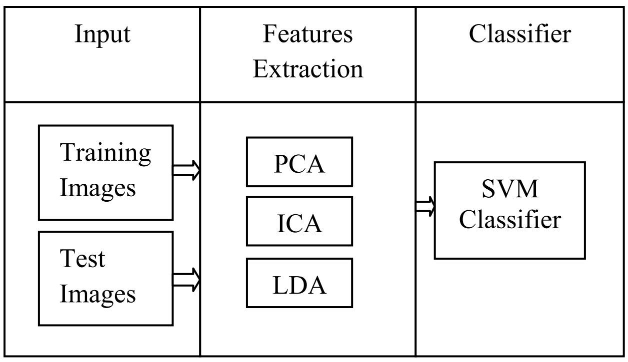


Fig 1: The face recognition system

The outline of the paper is as follows: Section 2 feature extraction and classification. In section 3 contains experimental results. Section 4 concludes the paper.

doi:10.4304/jetwi.6.1.64-68

II FEATURE EXTRACTION

Feature extraction involves several steps - dimensionality reduction, feature extraction and feature selection. We have a large features vector which considers the whole image that needs a reduction of dimension and selection the important features. Then these new features will be used for the training and testing of SVM classifier .In this paragraph we describe three techniques of extraction feature, Principal component analysis (PCA), independent component analysis (ICA) and linear discriminate analysis (LDA).

2.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a powerful tool for feature extraction as proposed by Turk and Pentland [13]. The main advantage of PCA is that it can reduce the dimension of the data without losing much information. Suppose there are N images Ii(i=l,2,---,N), each image is denoted as a column vector xi , and the dimension is M. The mean of the images is given by:

$$\bar{x} = \sum_{i=1}^{N} \frac{x_i}{N} \qquad (1)$$

the covariance matrix of images is given by

$$C = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{N}XX^T \qquad (2)$$

Where $X = [x_1 - \bar{x}, x_2 - \bar{x}, ..., x_N - \bar{x}]$ the projection space is made up of the eigenvectors which correspond to the significant eigenvalues when M>>N, the computational complexity is increased .we can use the singular value decomposition (SVD), theorem to simplify the computation .the matrix X, whose dimension is M*N and rank is N, can be decomposed as:

$$X = U\Lambda^{\frac{1}{2}}V^T \qquad (3)$$

$$U = X \vee \Lambda^{\frac{1}{2}} \qquad (4)$$

Where :

$\Lambda = diag[\lambda_1, \lambda_2, ..., \lambda_N], \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$, are the nonzero eigenvalues of XX^T and X^TX .

$U = [u_1, u_2, ..., u_M], V = [v_1, v_2, ..., v_N]$ are orthogonal matrices.

u_i is the eigenvector of XX^T , v_i is the eigenvector of X^TX and the λ_i is the corresponding eigenvalue.

u_i is calculated by following :

$$U_i = \frac{1}{\sqrt{\lambda_i}}Xv_i \qquad i = 1, 2, ..., N \qquad (5)$$

The p eigenvectors $U = [u_1, u_2, ..., u_p]$ $p \leq N$ corresponding to the p significant eigenvalues are selected to form the projection space and the sample feature is obtained by calculating.

2.2 Analyse discriminate linear (LDA)

LDA also known as Fisher's Discriminate Analysis, is another dimensionality reduction technique, it determines a subspace in which the between-class scatter (extra personal variability) is as large as possible, while the within-class scatter (intrapersonal variability) is kept constant. In this sense, the subspace obtained by LDA optimally discriminates the classes-faces.

We have a set of C-class and D-dimensional samples $\{x^{(1}, x^{(2}, ... x^{(N}\}$

N_1 of which belong to class w_1, N_2 to class w_2 and N_c to class w_c , In order to find a good discrimination of these classes we need to define a measure of separation, We define a measure of the within-class scatter by Eq. (6):

$$S_i = \sum_{x \in w_i}^{c}(x - \mu_i)(x - \mu_i)^T \qquad (6)$$

Where: $S_w = \sum_{i=1}^{c} S_i$ and $\mu_i = \frac{1}{N_i}\sum_{x \in w_i} x_i$

And the between-class scatter Eq. (7) becomes:

$$S_B = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T \qquad (7)$$

Where: $\mu = \frac{1}{N}\sum_{\forall x} x = \frac{1}{N}\sum_{i=1}^{C} N_i\mu_i$

Matrix $S_T = S_B + S_W$ is called the total scatter similarly, we define the mean vector and scatter matrices for the projected samples as:

$$\tilde{S}_W = \sum_{i=1}^{c}\sum_{y \in w_i}(y - \tilde{\mu}_i)(y - \tilde{\mu}_i)^T$$

$$\tilde{S}_B = \sum_{i=1}^{c} N_i(\tilde{\mu}_i - \tilde{\mu})(\tilde{\mu}_i - \tilde{\mu})^T$$

Where: $\tilde{\mu}_i = \frac{1}{N_i}\sum_{y \in w_i} y$, $\tilde{\mu} = \frac{1}{N}\sum_{\forall y} y$

From our derivation for the two-class problem, we can write: $\tilde{S}_B = W^T S_B W$ and $\tilde{S}_W = W^T S_W W$

Recall that we are looking for a projection that maximizes the ratio of between-class to within-class scatter. Since the projection is no longer a scalar (it has C−1 dimensions), we use the determinant of the scatter matrices to obtain a scalar objective function Eq. (8):

$$J(W) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \frac{W^T S_B W}{W^T S_W W} \qquad (8)$$

And we will seek the projection matrix W* that maximizes this ratio it can be shown that the optimal projection matrix W* is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem Eq. (9):

$$w^* = \left[w_1^* \middle| w_2^* \middle| .. w_{c-1}^*\right] = \arg\max \frac{|W^T S_B W|}{\|W^T S_W W\|} \Rightarrow (S_B - \lambda_i S_W) W_i^* \qquad (9)$$

S_B is the sum of C matrices of rank ≤1 and the mean vectors are constrained by : $\frac{1}{c}\sum_{i=1}^{C} \mu_i = \mu$

Therefore, S_B will be of rank (C−1) or less and this means that only (C−1) of the eigenvalues λ will be nonzero. The projections with maximum class separability information are the eigenvectors corresponding to the largest eigenvalues of $S_W^{-1} S_B$.

We seek (C−1) projections $\left[y_{1,2}, \ldots y_{c-1} \right]$ by means of $(c-1)$ projection vectors wi arranged by columns into a projection matrix $W = [w_1 \mid w_2 \mid \ldots \mid w_{c-1}] : y_i = w_i Tx \Rightarrow y = WTx$.

2.3 Independent Component Analysis (ICA)

The most common method for generating spatially localized features is to apply independent component analysis (ICA) to produce basis vectors that are statistically independent (not just linearly decorrelated , as with PCA) [14].it is an alternative to PCA which provides a more powerful data representation [15] and it's a discriminate analysis criterion, which can be used to enhance PCA.

ICA for face recognition has been proposed under two architecture by Barlett et. al. [16]. The architecture 1 aimed at finding a set of statistically independent basis images while the architecture 2 finds a factorial code. In this paper, the architecture 1 has been used. This process involves the following two initial steps :

1. The face images in the database are organized as a matrix X in which each row corresponds to an image.

2. The face database is processed to obtain a reduced dataset in order to reduce the computation efficiency of the ICA algorithm. The reduced dataset is obtained from the first m principal component (PC) eigenvectors of the image database. Hence the first step is applying PCA to determine the m PCs ,then the ICA algorithm is performed on the principal components using the mathematical procedure described in [17].

III Classification: Support Vector Machine (SVM)

SVMs (Support Vector Machines) are a useful technique for data classification and are still under intensive research [18],[19]. Although SVM is considered easier to use than Neural Networks, there are several kernels are being proposed by researchers, the four basic kernels as follow: linear, polynomial, sigmoid and radial basis function (RBF), We chose RBF kernel function for SVM classifier in our face recognition experiments Eq:10 which has fewer numerical difficulties [18] .

$$K\left(x_i, y_j\right) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \qquad \gamma \succ 0 \qquad (10)$$

γ is kernel parameter and parameterized using $\gamma = \dfrac{1}{2\sigma^2}$

3.1 Maximal Margin Hyperplanes

After we change the representation of the training examples by mapping the data to a features space F where the optimal separating hyper plane (OSH) is constructed Fig 2, we limited our study to the case of two-class discrimination [8] and we consider the training data S a set of l vectors features each vector has n dimension, where each point xi belongs to one of two classes identified by the label -1 or 1 Eq 11.

$$S = \left\{ \left(x_i, y_i \right) \middle| x_i \in R^n, y_i \in \{-1, 1\} \right\}_{i=1}^{l} \qquad (11)$$

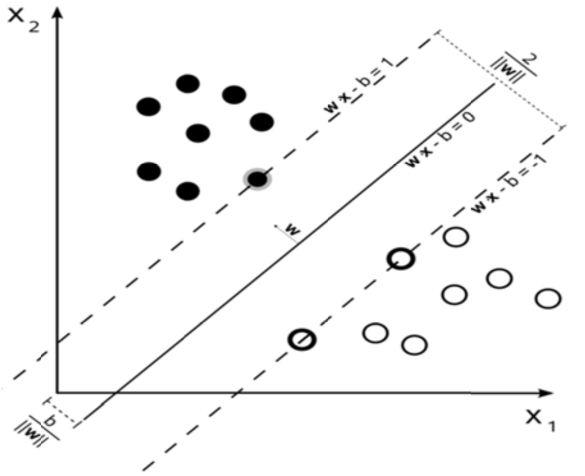


Fig 2: Maximum-margin hyperplane for SVM trained with samples from two classes.

We have solving a quadratic optimization problem with linear constraints that can be interpreted in terms of the Lagrange multipliers calculated by quadratic programming Eq: 12

$$\left\{ \begin{array}{c} \max(\alpha_i) : \tilde{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k\left(x_i, x_j\right) (\text{for any } i=1,..,n)? \quad \leq \alpha i \leq c \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{array} \right\} (12)$$

αi are the Lagrange multipliers parameters to be adjusted ,c is the penalty parameter of the classification error term it must be adjusted because the data are rarely completely separable, the xi are the training examples.

The solution of the optimization problem will be a vector $w \in F$, that can be written as a linear combination of the training inputs Eq: 13 $w = \sum \alpha_i y_i x_i \quad (13)$

(w,b) define the hyperplane $OSH = \{x : w.x + b = 0\}$ b is the bias.

We use the separating (OSH), once we have trained it on the training set, The (OSH) divides the R^n into two regions: one where $w.x_i + b \geq 0$ and one where $w.x_i + b \leq 0$. To use the maximal margin classifier, we determine on which side the test vector lies and assign the corresponding class label. Hence, the predicted class of a test point x is the output of the decision function Eq 14.

$$d(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i k\left(x_i, x\right) + b \right) \qquad (14)$$

$$K\left(x_i, y_j\right) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \qquad \gamma \succ 0$$

3.2 Multiclass Classification

SVM was originally designed for binary classification. Face recognition is a multi-class classification problem. There are two basic methods for face recognition with SVMs: one against-one and one-against-all. The one-against-one method is

Classification between each pair classes . The one-against-all is classification between each class and all the rest classes. In our experiments the one-against-all method was used for classification.

In real world problems we often have to deal with $n \geq 2$ classes. Our training set will consist of pairs (x_i, y_i), where $x_i \in R^n$ and $y_i \in \{1,...,n\}, i = 1..l$ for extending the two-class to the multiclass case this method will be described briefly below.

3.2.1 One vs. all approach

In the one-Vs-all approach n SVMs are trained. Each of the SVMs separates a single class from all remaining classes [20,21] ,A more recent comparison between several multi-class techniques [22] favors the one-vs-all approach because of its simplicity and excellent classification performance. Regarding the training effort, the one-vs-all approach is preferable over the one-vs-one approach since only n SVMs have to be trained compared to $n(n-1)/2$ SVMs in the pairwise approach (one-vs-one) [23], [24], [25] . The construction of a n-class classifier using two-class discrimination methods is usually done by the following procedure:

Construct n two-class decision functions $d_k(x), k = 1,..,n$ that separate examples of class k from the training points of all other classes,

$$d_k(x) = \begin{cases} +1 & \text{if x belongs to class k} \\ -1 & \text{otherwise} \end{cases}$$

In the face database of n individuals, 10 face images for everyone. 5 images among the 10 images of every one were taken to compose training samples and the rest 5 ones compose test samples.

Five images of first individual was taken and marked as positive samples, the all images of other training samples as negative samples. Both positive samples and negative samples were taken as input samples to train a SVM classifier to get corresponding support vectors and optimal hyperplane. The SVM was labeled as SVM1. In turn we can get the SVM for every individual and labeled as SVM1, … , SVMn respectively.

The n SVMs can divide the samples into n classes. When a test sample was in turn inputted to every SVM, there would be several cases:

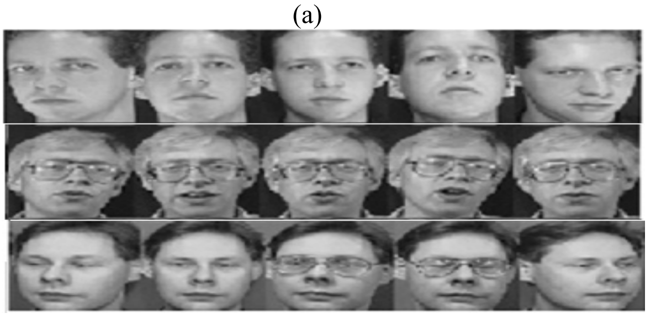
- If the sample was decided to be positive by SVMi and to be negative by others SVMs at the same time, then the sample was classified as class i.
- If the sample was decided to be negative by several SVMs synchronously and to be positive by other SVMs, then the classification was false.
- If the sample was decided to be negative by all SVMs synchronously, then the sample was decided not belonging to the face database.

IV. EXPERIMENTATION AND RESULTS

Our experiments were performed on two face databases, The ATT Face Database [1] and the Indian Face Database (IFD) [2] the ATT database contains images with very small changes in orientation of images for each subject involved, while the IFD contains a set of 10 images for each subject where each image is oriented in a different angle compared to the other.

These two databases both contains 10 classes, each class have 5 images for training and 5 images for testing Fig 3 and Fig 4. We use these Databases for comparison of different face recognition algorithms such as PCA+SVM, LDA+SVM and ICA+SVM. We extract different features from a training set and testing set using PCA, LDA, ICA methods. Using these feature we trained the classifier SVM and find the accuracy of the three methods, the recognition rates of the three methods PCA+SVM, LDA+SVM, ICA+SVM were shown as Fig. 5.

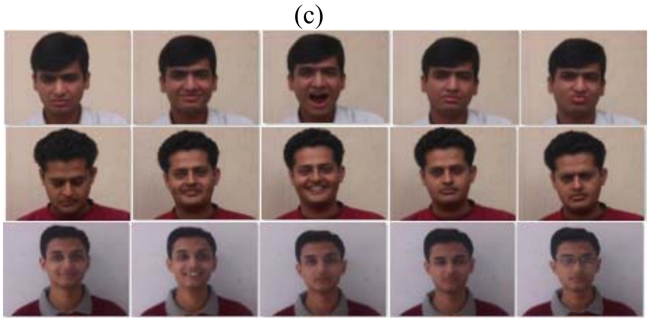
(a)



(b)

Fig 3: Examples of (a) training and (b) test images of (ATT) Face Database

(c)



(d)

Fig 4: Examples of (c) training and (d) test images of (IFD) Face Database

The comparison is done on the basis of rate of recognition accuracy. Comparative results obtained by testing the PCA+SVM, LDA+SVM, ICA+SVM algorithms on both the IFD and the ATT databases Fig.5.

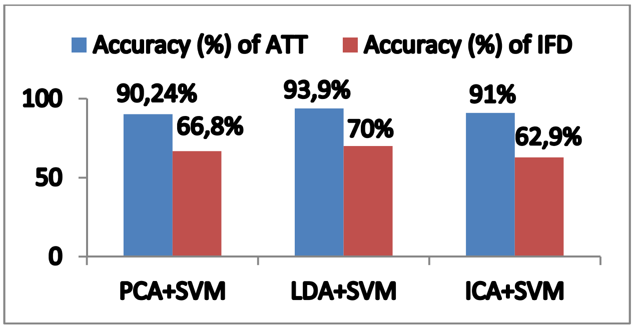


Fig 5: Comparative of the combination algorithms PCA+SVM,LDA+SVM,ICA+SVM On the basis Of recognition accuracy

It is observed that recognition rate of the method LDA+SVM is 93.9% obtained on ATT face database and 70% on IFD face database it is the higher as compare to PCA+SVM and ICA+SVM methods for both IFD and ATT databases.

CONCLUSION

We presented a face recognition method based on SVM classifier combined with LDA feature extraction. We implemented experiments on IFD and ATT face database. First, LDA for dimension reduction and SVM for classification. The experimental results showed that LDA+SVM method had a higher recognition rate than the other two methods for face recognition.

REFERENCES

[1] AT&T Laboratories Cambridge,http://www.cl.cam.ac.uk/Research/DTG/attarchive:pub/data/att_faces.tar.Z.
[2] Vidit Jain and Amitabha Mukherjee, The Indian Face Database.http://vis-www.cs.umass.edu/~vidit/IndianFaceDatabase/, 2002
[3] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. Journal of the Optical Society of America A - Optics, Image Science and Vision, 4(3):519–524, March 1987.
[4] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(1):103–108, 1990.
[5] M. Turk and A. P. Pentland, “Eigenfaces for recognition”, Journal of Cognitive Neuroscience, vol. 3, no. 1,pp. 71–86, 1991.
[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: recognition using class specific linear projection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7,pp. 711–720, 1997.
[7] M. Bartlett, J. Movellan, and T. Sejnowski. Face recognition by independent component analysis. IEEE Trans. on Neural Networks,13(6):1450–1464, November 2002.
[8] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, New York, 1998.
[9] Vapnik V. “Statistical Learning Theory”, Wiley, New York, 1998.
[10] G. Paliouras, V. Karkaletsis, and C.D. Spyropoulos (Eds.), “Support Vector Machines: Theory and Applications,” ACAI ’ 99, LNAI 2049,pp. 249-257, 2001.
[11] Nello Cristianini, John Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, 2000.
[12] Vladimir N. Vapnik, The Nature of Statistical Learning Theory,Springer, New York, 2000.
[13] M. A. Turk and A. P. Pentland, "Face Recognition Using Eigenfaces," in IEEE CVPR), 1991,pp. 586-591.
[14] M. S. Bartlett, Face Image Analysis by Unsupervised Learning: Kluwer Academic, 2001.
[15] C. Liu and H. Wechsler. Comparative assessment of independent component analysis (ica) for face recognition. In Proc. of the Second In-ternational Conference on Audio- and Video-based Biometric Person Authentication, AVBPA’99, Washington D.C., USA, March 1999.
[16] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, Face Recognition by Independent Component Analysis, IEEE Trans. on Neural Networks, Vol. 13, No. 6, November 2002, pp. 1450-1464
[17] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, Face Recognition by Independent Component Analysis, IEEE Trans. on Neural Networks, Vol. 13, No. 6, November 2002, pp. 1450-1464
[18] Vapnik, V., 1995: The Nature of Statistical Learning Theory. Springer, N.Y.
[19] Vapnik, V., 1998: Statistical Learning Theory. Wiley, N.Y.
[20] C. Cortes, V. Vapnik, Support vector networks, Mach. Learning 20 (1995) 1–25.
[21] B. Sch€olkopf, C. Burges, V. Vapnik, Extracting support data for a given task, in: U. Fayyad, R.Uthurusamy (Eds.), Proc. First Int. Conf. on Knowledge Discovery and Data Mining, AAAI Press,Menlo Park, CA, 1995.
[22] R. Rifkin, Everything old is new again: a fresh look at historical approaches in machine learning,Ph.D. thesis, M.I.T., 2002.
[23] M. Pontil, A. Verri, Support vector machines for 3-d object recognition, IEEE Trans. Pattern Anal.Mach. Intell. (1998) 637–646.
[24] G. Guodong, S. Li, C. Kapluk, Face recognition by support vector machines, in: Proc. IEEE Int.Conf. on Automatic Face and Gesture Recognition, 2000, pp. 196–201.
[25] Platt J., N. Cristianini, and J. Shawe-Taylor. Large margin dags for multiclass classification. In Advances in Neural Information Processing Systems 12, pages 547-553. MIT Press, 2000.