BOOSTING CHROMATIC INFORMATION FOR FACE RECOGNITION

T. Ganapathi, Student Member,IEEE*, K.N. Plataniotis,* Senior Member,IEEE*, Y.M. Ro,* Senior Member, IEEE*

The Edward S Rogers Sr Department of Electrical and Computer Engineering
University of Toronto
10 King's College Road, Toronto, ON M5S 3G4
email: {tejas,kostas,yro}@comm.utoronto.ca

ABSTRACT

In this paper, chromatic information is integrated with an adaboost learner to address non linearities in face patterns and illumination variations in training databases for face recognition (FR). An LDA based learner is boosted and the integrated framework is tested on a large database of images having severe pose and illumination variations. The increased dimensionality of color induces a small sample size problem when used with an LDA based learner. The integrated framework is tested on a number of learning scenarios in order to examine this effect. Experimental results show that integrating color into the boosting framework produces a high performing FR system for a range of learning scenarios.

Index Terms— Color Face Recognition, Adaptive Boosting, Linear Discriminant Analysis, Small Sample Size

1. INTRODUCTION

Face Recognition (FR) has applications ranging from monitoring and surveillance, human computer interaction to identity authentication. Among the FR methods proposed, appearance based approaches are shown to be the most successful[1, 2, 3]. They focus on low dimensional statistical feature extraction and operate on the face as a 2-d pattern, avoiding 3-d modeling and landmark detection. Of all classes of appearance based approaches, supervised learning methods based on LDA have shown promising results [2, 1, 3, 4].

Color features lead to a better recognition performance than gray scale information alone[5, 6] and it makes object recognition robust against illumination variations. A discriminative feature space representing chromatic and intensity information created using a supervised learner such as LDA is hence expected to enhance the FR system combining the advantages of chromatic features and supervised learning. However, LDA based methods are susceptible to the small sample size problem [1, 2, 7, 3] faced in high dimensional pattern recognition tasks like FR which becomes more significant when multi-spectral (or color) images with higher dimension are used. In addition, the performance of LDA based methods deteriorate when face patterns are subject to severe nonlinearities: variations in expression, pose, and illumination. To take these into account while training the system, learning models like LDA should be replaced by either globally nonlinear models, or by a linear combination of locally linear models. The latter are advantageous over non linear methods as they are less likely to over fit and do not require optimization of many parameters [3].

In this paper, the non linearities in face patterns and illumination variations are addressed by combining the advantages of chromatic features with ensemble learning. The B-JD-LDA [3] method for ensemble learning which is based on boosting a learner consisting of JD-LDA feature extractor [1] (a variant of direct LDA)and a linear classifier is chosen as our ensemble learner, as it combines the advantages of adaptive boosting in addressing non-linearities, and direct LDA in addressing degenerate scatter matrices. Since the small sample size problem is worsened when color inputs are used, we have tested this integrated framework on a range of small sample size scenarios, to examine the effect of small sample size parameter on both color images and boosting framework. Our evaluation database consists of 4760 images and our results are reported in Section 4.

2. REPRESENTATION OF COLOR INFORMATION

Let S_i be the i^{th} 2-d image in a set of images, with spatial dimensions, $J = I_W \times I_H$ and K spectral planes. Each spectral plane has a spectral depth of 8 bits whose values $\in I$ and lie in [0,255]; therefore S_i has a spectral depth of $K \times 8$ bits. S_i is represented as a column vector, x_i by the following steps: 1) each spectral plane is converted to a column vector 2) the column vectors of each spectral plane are concatenated. In order to form a column vector for the m^{th} spectral plane of the i^{th} image, where $1 \leq m \leq K$, the 8 bit values of that spectral plane are ordered lexicographically into a column vector, s_{im}, where $s_{im} \in R^{D_m \times 1}$. The dimensionality of s_{im}, D_m, is dependent on the sampling nature of s_{im} and

*On research leave from Information and Communications University, Daejon, Korea

CCECE/CCGEI May 5-7 2008 Niagara Falls. Canada
978-1-4244-1643-1/08/$25.00

the spatial dimension of S_i. For example, if S_i is converted to the YCbCr 4:2:0 color space used in MPEG-1 standards, then $K = 3$, and sub sampling is performed on Cb and Cr spectral planes while forming their column vectors. Therefore, $D_Y : D_{Cb} : D_{Cr} = 4 : 1 : 1$, and their values will be of the form $D_m = J/\mu$, where μ is the scaling factor of that particular spectral plane. In the case of YCbCr, $\mu = 0.25$ for the Cb % Cr and $\mu = 1$ for Y. After forming the column vectors, $s_{i1}, ..., s_{iK}$, x_i is formed by, $x_i = [s_{i1}^T s_{i2}^T ... s_{iK}^T]^T$. The Dimension of x_i is $d = \sum_{m=1}^{K} D_m$.

Small Sample Size Problem: Supervised learning FR methods based on LDA typically face a small sample size problem as the number of samples per subject, L available for training is very small (≥ 10) compared to the dimensionality of the column vectors in the training set, Z (of the order of $\approx 10^4$). This makes the estimation of the within class scatter matrix, S_W a highly ill posed problem. For inputs with multiple spectral planes ($K \geq 2$), the dimensionality of d is increased by factor K, thus making the estimation of S_W more ill posed. For example, if x_is are gray scale and have a dimensionality of $d = 150 \times 130 = 19500$ and the number of training samples per subject, $L = 2$ for all subjects, the small sample size becomes technically worse if x_i was color (without sub sampling) as d would be increased 3 times to 58500, while L would remain the same.

3. ADAPTIVE BOOSTING FRAMEWORK

In order to address the non linearities in the distribution of face patterns, we have used an ensemble based learning method based on the adaptive boosting framework. The B-JD-LDA [3] method of adaptive boosting is chosen owing to its demonstrated capability in addressing a large database containing non-linearities in the form of pose and expression variations on gray scale images. The individual learner in the B-JD-LDA consists of a JD-LDA[1] feature extractor and a Nearest Center Classifier and is referred to as a *g-Classifier*. The JD-LDA feature extractor is a variant of the direct LDA[2] and uses a modified fisher's criterion which finds the basis vectors which maximize the ratio of the between class scatter matrix to total scatter matrix, thus avoiding the inversion of a singular S_W. A new *g-Classifier* is formed in each subsequent iteration based on the *feedback* from the previous learner in the form of the updated parameters, which depend on the error in hard to classify samples and hard to classify subjects of the previous iteration. The final classifier is a weighted sum of all *g-Classifiers*. A general diagram depicting adaptive boosting framework is given in Figure 1.

For optimal performance of the boosting method, the individual learners of the B-JD-LDA should have a low mutual dependence with each other and a low generalization error on the training set. The boosting method does not perform better over iterations if either the individual *g-Classifiers* are too strong, i.e., have a high mutual dependence, or they are too

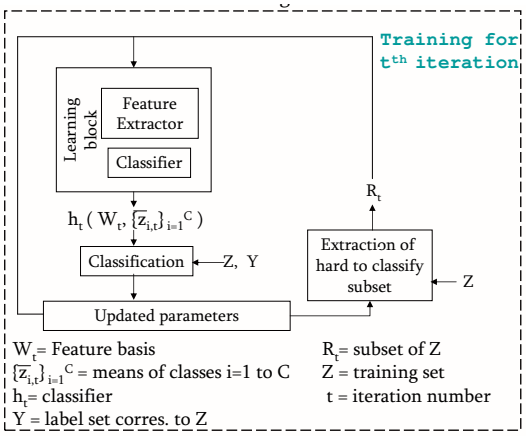

Fig. 1. Ada-Boost Framework: Training for the tthiteration

weak so as to produce a very high generalization error. The samples per subject available for training each *g-Classifier*,r is therefore used as a parameter for adjusting the weakness of the *g-Classifiers*. This trade off is best achieved by using a *g-Classifier* with optimal weakness. The weakness of the *g-Classifier* is described using a quantity called the *Learning Difficulty Degree* (LDD) which is the ratio of r to C. The trade off between weak *g-Classifiers* and low generalization error, is achieved by choosing an optimal $r*$ which will differ for each learning scenario and $\in [2,L]$. When the images in Z are color, d increased by a factor of K. Increased dimensionality induces a small sample size problem in training the *g-Classifiers*, which could increase the generalization error on Z. We are interested in the effect of color images on the $r*$ of the individual *g-Classifiers* and the booster.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

For our experiments, we have chosen a subset of the color PIE database [8] for evaluation, which demonstrates the high complexity of the face pattern distribution. Since we want to evaluate the effect that color has on the FR system, we have considered the RGB and YCbCr 4:4:4 color space transformations along with their gray scale counterparts: R and Y. We have chosen the RGB color space because it is the form in which color images are usually displayed, and the YCbCr as it is used in MPEG compression standards. All experiments are performed using both the B-JD-LDA and JD-LDA methods for both pairs of color space and gray scale transformations to examine the improvement obtained by boosting the learner.

The CMU PIE database consists of 70 subjects under varying conditions of pose & lighting, pose & illumination and pose & expression. Since the ground truth data (coordinate information of the centers of the eyes and nose tips) for

only 68/70 subjects are provided, we have used all 68 subjects for experiments. We have included 7 poses(horizontal and vertical variations up to 45^o) and 10 illumination conditions (caused by varying positions of the camera flash in a room with zero background illumination). Following the PIE's naming rule, pose group [05,07,09,11,27,29,37] and flash numbers [2,4,6,10,12-14,16,18,19] are chosen. Details of the evaluation subset used for experimentation, D are: 1) No. of samples/subject: 70. 2)Total number of images in the evaluation database: 68 subjects$\times$70=4760.

The images in the databases contain irrelevant information along with the face. To avoid incorrect evaluations, following are the sequence of processing steps [3, 9]: 1) The raw color image, of spatial dimension 480$\times$640 pixels is translated, rotated and scaled to 150×130, so that the centers of the eyes are placed on definite pixels, the distance between the eye centers is 70 pixels and eye centers are placed on the 45th row to maintain the photometric proportion of the face. 2) Standard mask is applied to this image of above spatial dimension to remove the non-face portions, 3) The required color space/ gray scale transformation is applied and the column vector is formed. The resolution of 150×130 was chosen as it is used in surveillance applications. Following standard FR practices, D is divided into two sets: the gallery set on which training is performed, Z and the probe set, Q which contains the images of unknown identity, such that $D = Z + Q$, and $Z \cap Q = \emptyset$. L images per subject from D comprise Z, while the remaining $70 - L$ images per subject constitute Q. Experiments have been conducted for $L \in \{3, 4, 5, 6, 7, 10, 13, 16\}$ in order to examine effect of increased dimensionality caused by color and the performance of the integrated framework on a range of small sample size scenarios. The images of each subject chosen to comprise Z are ensured to be of different poses (if $L \leq 7$) and different illuminations, so that all 7 poses and 10 illuminations are represented by the 68 subjects. The B-JD-LDA and JD-LDA trained on Z and evaluated on Q to produce a Classification Error Rate (CER) which is the ratio of the number of wrong identifications to the number of probe images. The Nearest Center Classifier used is based on Euclidean distance. To incorporate into the ada-boost framework, the classification score should have values in [0,1] and is calculated as follows,

$$dist(z, i, \Psi_t, \bar{z}_{i,t}) = \frac{dist_{max} - dist_{z,i}}{dist_{max} - dist_{min}} \qquad (1)$$

where $dist_{z,i} = \left\| \Psi^T (z - \bar{z}_i) \right\|$, $dist_{max} = max(\{dist_{z,i}\}_{i=1}^{C})$ and $dist_{min} = min(\{dist_{z,i}\}_{i=1}^{C})$; z is the column vector of the training image, i is the class number, Ψ_t is the JD-LDA feature basis at iteration t, $\bar{z}_{i,t}$ is the mean of class i at iteration t. The results reported are at an average `10`runs; each run is executed on a gallery and probe partition. The assumption here is that the *same color/ gray scale transformation is used in both testing and training sets*; that is the FR user knows the transformation used for training the system.

4.2. Comparison of FR performances

In this section, the effects of using chromatic information on the FR system and boosting the JD-LDA learner is presented. In order to examine these effects, we have used the following quantities- ξ_B^* and ξ_J^*: best improvements obtained by color space transformation over its gray scale counterpart for the B-JD-LDA and JD-LDA respectively, and δ^*: best improvement obtained by using B-JD-LDA over JD-LDA. Negative quantities imply performance improvements.

Color inputs improve performance of the FR system over their gray scale counterparts for all the learning scenarios and algorithms examined. YCbCr color space shows a better improvement over its gray scale counterpart than the RGB color space. From table 1, $|\xi_B^*|$ and $|\xi_J^*|$ are significantly greater for the YCbCr & Y pair of transformations compared to the RGB & R pair, which implies YCbCr is a better color space for FR. For both pairs of transformations $|\xi_B^*|$ and $|\xi_J^*|$ are highest when $L = 4$ and monotonically decrease as L increases. $|\xi_B^*|$ and $|\xi_J^*|$ are marginally lower for the $L = 3$ case and this observation can be attributed to the effect of the increased dimensionality of color on the small sample size problem. However, their values are still significantly high to use chromatic information over gray scale. Another trend observed is, $|\xi_J^*|$ reduces less rapidly as L increases than $|\xi_B^*|$, i.e., as the learning scenario becomes less *hard*, and the ensemble learner becomes stronger, the improvement offered by chromatic information is reduced.

The B-JD-LDA has a better performance than JD-LDA method of FR for all examined cases. $|\delta^*|$ is larger when the size of the training database is large, i.e., >4. The value $|\delta^*|$ is not significant for the case when $L = 3$, however is over 6% for all cases when L>4. $|\delta^*|$ is almost equally high for both color and gray scale transformations $\forall L$, i.e., the boosting the learner produces improvement for both color and gray scale inputs. As proved in earlier literature which utilizes the B-JD-LDA [3] for face recognition, $r*$ should not be too high or too low. However no trend in the shift of $r*$ was observed when color transformations were used instead of their gray scale counterparts.

Chromatic information offers a significant improvement to the FR system in small sample size scenarios, while boosting the learner does not significantly improve the system. As the value of L is increased, the improvement provided by color information reduces. Boosting the learner improves the performance of the individual learner significantly in all cases where the size of the training database is reasonably large, i.e., $L \geq 4 \& |D| > 272$ images. The experimental results show that integrating color into the boosting framework could significantly improve the performance of the FR system when $L \approx 4 - 10$ for medium sized databases.

Table 1. Results obtained with B-JD-LDA & JD-LDA using color & gray scale transformations in different learning scenarios

		Gray scale transformations				Color Space Transformations					
		R		Y		RGB			YCbCr		
Samples/ Subject	r	B-JD-LDA(T*) CER %	JD-LDA(M*) CER%, δ	B-JD-LDA(T*) CER %	JD-LDA(M*) CER%, δ	B-JD-LDA(T*) CER %	JD-LDA(M*) CER%, δ	ξ %	B-JD-LDA(T*) CER %	JD-LDA(M*) CER%, δ	ξ %
3	2	65.86(6)	CER= 56.3191(46)	64.45(4)	CER=56.3586(46)	60.14(11)	CER= 52.7283(46)	ξ^*_B -5.1	59.27(1)	CER= 46.385(46)	ξ^*_B -8.28
	3	54.97(5)	δ^*=−1.349 (r^*=3)	54.17(7)	δ^*=-2.189 (r^*=3)	49.87(21)	δ^*=-2.8583 (r^*=3)	ξ^*_J -3.59	45.89(3)	δ^*=-0.495 (r^*=3)	ξ^*_J -9.974
4	2	67.9(5)	CER=46.8326(35)	64.12(5)	CER=46.884(35)	60.72(5)	CER= 41.9546(35)	ξ^*_B -5.79	61.33(2)	CER= 36.7647(35)	ξ^*_B -8.91
	3	39.95(30)	δ^*=−6.882 (r^*=3)	38.62(23)	δ^*=-8.264 (r^*=3)	34.16(30)	δ^*=-7.7946 (r^*=3)	ξ^*_J -4.88	29.71(24)	δ^*=-7.054 (r^*=3)	ξ^*_J -10.12
5	2	67.94(1)		65.37(4)		65.15(3)			62.09(2)		
	3	35.36(40)	CER=39.5989(33)	32.38(40)	CER= 39.2534(33)	29.71(38)	CER= 35.0309(33)	ξ^*_B -5.14	25.15(36)	CER= 29.2464(33)	ξ^*_B -8.55
	4	31.20(37)	δ^*=−8.3989 (r^*=4)	29.12(34)	δ^*=-10.133(r^*=4)	26.06(34)	δ^*=-8.9709 (r^*=4)	ξ^*_J -4.57	20.57(22)	δ^*=-8.6764 (r^*=4)	ξ^*_J -10.01
6	3	33.86(40)		28.54(40)		26.79(40)			29.64(7)		
	4	25.17(39)	CER= 33.3203(30)	23.84(27)	CER= 33.596(30)	20.78(40)	CER = 28.8166(30)	ξ^*_B -4.06	16.76(36)	CER= 23.8074(30)	ξ^*_B -6.24
	5	23.67(37)	δ^*=−9.6503 (r^*=5)	22.47(23)	δ^*=-11.126 (r^*=5)	19.61(33)	δ^*=-9.2066 (r^*=5)	ξ^*_J -4.5	16.23(17)	δ^*= -7.5774 (r^*=4)	ξ^*_J -9.789
7	3	32.64(36)		23.92(40)		21.87(37)			31.18(4)		
	4	19.78(39)		17.32(37)		14.35(39)			11.07(37)		
	5	16.71(40)	CER= 27.1761(30)	15.46(31)	CER= 27.0003(30)	12.33(29)	CER= 22.1548(30)	ξ^*_B -3.88	9.37(30)	CER= 17.411(30)	ξ^*_B -5.76
	6	16.04(40)	δ^*=-11.136 (r^*=6)	15.13(21)	δ^*=-11.8703 (r^*=6)	12.16(38)	δ^*=-9.9948 (r^*=6)	ξ^*_J -5.02	9.76(13)	δ^*=-8.041 (r^*=5)	ξ^*_J -9.589
10	5	13.73(38)		12.56(40)		10.62(39)			7.21(40)		
	6	12.50(39)		11.57(30)		9.70(34)			6.28(40)		
	7	11.87(39)	CER=20.9657(30)	11.37(35)	CER =21.5931(30)	9.42(39)	CER= 17.3676(30)	ξ^*_B -2.45	6.04(28)	CER= 13.4559(30)	ξ^*_B -5.33
	9	13.79(22)	δ^*=-8.825 (r^*=7)	13.80(8)	δ^*=-7.4069 (r^*=7)	11.39(15)	δ^*= -7.9476 (r^*=7)	ξ^*_J -3.6	8.04(6)	δ^*=-7.4159 (r^*=7)	ξ^*_J -8.137
13	5	16.67(36)		14.95(39)		12(40)			9.13(39)		
	7	13.28(40)		12.45(39)		9.36(40)			7.57(38)		
	9	12.63(31)	CER= 21.9401(30)	12.24(36)	CER =22.549(30)	9.66(40)	CER= 19.1744(30)	ξ^*_B -3.27	8.50(11)	CER= 14.6646(30)	ξ^*_B -4.67
	11	15.07(14)	δ^*=-9.31 (r^*=9)	14.14(10)	δ^*=-9.451 (r^*=9)	12.16(40	δ^*=-9.8144 (r^*=8)	ξ^*_J -2.77	9.73(7)	δ^*=-7.0946 (r^*=8)	ξ^*_J -7.884
16	5	15.1(40)		13.27(40)		9.91(40)			8.42(40)		
	7	10.86(40)		10.43(39)		7.19(40)			6.06(35)		
	9	9.59(35)		9.8(40)		6.57(37)			5.56(26)		
	11	9.71(39)		10.09(29)		7.43(36)			6.69(11)		
	13	11.49(24)	CER= 17.963(30)	11.43(15)	CER = 18.6819(30)	9.16(24)	CER= 15.2614(30)	ξ^*_B -3.02	7.41(10)	CER= 11.6285(30)	ξ^*_B -4.24
	15	13.09(17)	δ^*=-8.373 (r^*=9)	12.86(20)	δ^*=-8.8819(r^*=9)	10.59(12)	δ^*= -8.6914 (r^*=9)	ξ^*_J -2.7	8.02(7)	δ^*=-6.0685 (r^*=9)	ξ^*_J -7.053

All Classification Error Rate (CER) are expressed as a percentage
The CERs reported for the B-JD-LDA are the minimum over 40 ada-boost iterations, T denotes the iteration number at which this minimum was achieved*
no of JD-LDA features used =30 for all boosting experiments (JD-LDA is the feature extractor in the g-Classifier)
The CERs reported for JD-LDA are for the best found feature number for that learning scenario, where M is the most optimal number of features*

5. REFERENCES

[1] J.Lu, K.N.Plataniotis, and A.N.Venetsanopoulos, "Face recognition using lda-based algorithms," *IEEE Trans. on Neural Networks*, vol. 14, no. 1, pp. 195–200, Jan 2003.

[2] H.Yu and J. Yang, "A direct lda algorithm for high- dimensional data with application to face recognition," *Pattern Recognition*, vol. 34, pp. 2067–2070, 2001.

[3] J.Lu, K.N.Plataniotis, A.N.Venetsanopoulos, and S.Z. Li, "Ensemble-based discriminant learning with boosting for face recognition," *IEEE Trans. on Neural Networks*, vol. 17, no. 1, pp. 166–178, Jan. 2006.

[4] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *proc. of 4th European Conf. on Computer Vision*, pp. 45–58, Apr 1996.

[5] Shih Peichung and Liu Chengjun, "Comparative assessment of content based face image retrieval in different colour spaces," *Intl. Jnl. of Pattern Recognition*, vol. 19, no. 7, pp. 873–893, 2005.

[6] Shih Peichung and Liu Chengjun, "Improving the face recognition grand challenge baseline performance using color configurations across color spaces," *2006 IEEE Intl. Conf on Image Processing*, pp. 1001–1004, 8-11 Oct. 2006.

[7] S.J.Raudys and A.K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, Mar 1991.

[8] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination and expression database," in *Proc. of the 5th Intl. Conf on Automatic Face and Gesture Recognition*, Washington, D.C., 2002.

[9] P.J. Philips, H. Moon, S.A. Rizvi, and P. Rauss, "The feret evaluation methodology for face recognition algorithms," *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.