ISSN 2047-3338

2D-LDA based Online Handwritten Kannada Character Recognition

M. Mahadeva[1] Prasad and M. Sukumar[2]

[1]Department of Studies in Electronics University of Mysore, Post Graduate Centre, Hassan, India
[2]Department of Instrumentation Technology, S. J. College of Engineering, Manasagangothri Campus, Mysore, India
[1]prasada9@gmail.com, [2]m_sukumar@rediffmail.com

***Abstract*– Matrix based LDA (2D-LDA) is gaining importance as it overcomes the singularity problem in small sample size problem and also achieves better computational efficiency when compared to classical 1D-LDA. In 2D-LDA data is represented in matrix form rather than vector form before dimensionality reduction. In the present work, recognition performance of 2D-LDA for different feature combinations is compared in recognizing online handwritten Kannada Characters. Kannada basic character Set, popularly known as Kannada Varnamale is subjected to experimentation. Different feature combinations extracted from the size normalized characters are fed to 2D-LDA for dimensionality reduction and nearest neighbor classifier is used for classification. Writer independent experiments are carried out with 3750 samples for training and 1550 samples for testing. Among different feature combinations, second derivative feature and estimate feature combination has given a best average recognition accuracy of 88.3 %.**

***Index Terms*– Kannada Characters, Kannada Varnamale Recognition, Online Handwriting and 2D-LDA**

I. INTRODUCTION

MANY subspace algorithms used for dimensionality reduction and feature extraction are reported in the literature. Among these, principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2] are those which are reported in the beginning and are widely used for subspace feature extraction and dimensionality reduction. Now these are also called as classical PCA and classical LDA due to the existence of different variants of these basic algorithms. While PCA seeks a projection that best represents the data in the lease-squares sense, LDA seeks a projection that best separates the data in a least-squares sense [3]. Depending on the application and the nature available data either PCA or LDA are used. In the field of image classification, where the available class specific data samples are limited and the dimension of each sample is huge, PCA suffers from higher covariance matrix computation cost and LDA suffers from the singularity problem. To overcome these drawbacks, variants of PCA and LDA like PCA+LDA [2], Orthogonal LDA [4], regularized LDA [5], etc., have been proposed in the literature.

In all the above mentioned subspace algorithms, data is represented in column vector format. So, before applying these subspace tools, all the extracted features from a sample are vertically cascaded to form a column vector. In the case of image data classification, this leads to higher computation cost due to lengthy column vector and also singularity problem due to limited sample size. To overcome this, matrix-based data representation or two-dimensional data representation algorithms have evolved. As a result of this 2D-PCA [6] and 2D-LDA [7] have been proposed to directly work on image data.

In the present work, to better understand 2D-LDA, experiments are carried out to recognize online handwritten Kannada basic characters. This basic character set comprising of 50 characters is popularly known as Kannada varnamele. Combinations of different features extracted from the characters are subjected to experimentation. Among the feature combi-nations, second derivative feature and estimate feature combination has given a best average recognition accuracy of 88.3%.

II. AN OVERVIEW OF 2D-LDA

The explanation of basic principle behind 2D-LDA and its applications to the field of pattern classification can be found in many literatures. In order to give completeness to the paper, an overview of 2D-LDA is explained here:

Let $\{X_{1,} X_{2,}X_{3,} \ldots\ldots., X_m\}$ be a set of m training data matrices of C classes, where, X_i $(i = 1, 2, \ldots., C)$ are data matrices in $rowxcol$ -dimensional space and t_i $(i = 1, 2, \ldots\ldots, C)$ be the number of training samples of i^{th} class. Let $\mu_i(i = 1, 2, \ldots\ldots\ldots.., C)$ and μ_o be the i^{th} class means and the global mean defined as:
$\mu_i = \frac{1}{t_i} \sum_{k=1}^{t_i} X_k$, $i = 1, 2, \ldots\ldots.., C$ and $\mu_o = \sum_{k=1}^{C} \frac{t_k}{N} \mu_k$. Where, $N = \sum_{k=1}^{C} t_k$ is the sum of the training samples of all the classes. Let $\mathcal{L} \otimes \mathcal{R}$ be a tensor product of two spaces $\mathcal{L}\ and\ \mathcal{R}$, which are spanned by $\{u_{1,}u_{2,}u_{3,} \ldots\ldots\ldots., u_r\}\ and\ \{v_{1,}v_{2,}v_{3,} \ldots\ldots\ldots.., v_c\}$ respectively and the corresponding matrices are, $\mathcal{L} = [u_{1,}u_{2,}u_{3,} \ldots, u_r]\ and\ \mathcal{R} = [v_{1,}v_{2,}v_{3,} \ldots, v_c]$.
If S_b^2 & S_w^2 are between-class and within-class matrices of two-

dimensional data and noting that the similarity metric between matrices is the Frobenius norm, these matrices can be defined as:

$$S_b^2 = \sum_{i=1}^{C} t_i \|\mathcal{L}^T(\mu_i - \mu_o)\mathcal{R}\|_F^2$$

and

$$S_w^2 = \sum_{p=1}^{C}\sum_{q=1}^{t_i} \left\|\mathcal{L}^T\left(X_p^q - \mu_i\right)\mathcal{R}\right\|_F^2$$

Since, $\|XX^T\|_F^2 = trace(XX^T)$, the above equations can be written as:

$$\widetilde{S_b^2} = trace\left(\sum_{i=1}^{C} t_i\ \mathcal{L}^T(\mu_i - \mu_o)\mathcal{R}\mathcal{R}^T(\mu_i - \mu_o)^T\mathcal{L}\right)$$

and

$$\widetilde{S_w^2} = trace\left(\sum_{p=1}^{C}\sum_{q=1}^{t_i} \mathcal{L}^T\left(X_p^q - \mu_o\right)\mathcal{R}\mathcal{R}^T\left(X_p^q - \mu_o\right)^T\mathcal{L}\right)$$

Now the feature extraction is to find the optimal matrices $\mathcal{L}\ and\ \mathcal{R}$ such that the original class structure is preserved by transforming original data matrix X_i into lower-dimensional space, rxc by $Y_i = \mathcal{L}^T X_i \mathcal{R}$ and the corresponding criterion $J(\mathcal{L},\mathcal{R})$, defined below is to maximizing $\widetilde{S_b^2}$ and minimizing $\widetilde{S_w^2}$.

$$J(\mathcal{L},\mathcal{R}) = \frac{\widetilde{S_b^2}}{\widetilde{S_w^2}}.$$

Since simultaneous computation of $\mathcal{L}\ and\ \mathcal{R}$ is not possible, two optimization functions [8] are defined to compute $\mathcal{L}\ and\ \mathcal{R}$.

By fixing $\mathcal{R}$, $\mathcal{L}$ is computed by solving the optimization function,

$$J_1(\mathcal{L}) = mxtrace\left((\mathcal{L}^T\ S_w^{2\mathcal{R}}\mathcal{L})^{-1}(\mathcal{L}^T\ S_b^{2\mathcal{R}}\mathcal{L})\right)$$

where,

$$S_b^{2\mathcal{R}} = \sum_{i=1}^{C} t_i(\mu_i - \mu_o)\mathcal{R}\mathcal{R}^T(\mu_i - \mu_o)^T$$

and

$$S_w^{2\mathcal{R}} = \sum_{p=1}^{C}\sum_{q=1}^{t_i}\left(X_p^q - \mu_o\right)\mathcal{R}\mathcal{R}^T\left(X_p^q - \mu_o\right)^T$$

Similarly, by fixing $\mathcal{L}$, $\mathcal{R}$ is computed by solving the optimization function,

$$J_2(\mathcal{R}) = mxtrace\left((\mathcal{R}^T\ S_w^{2\mathcal{L}}\mathcal{R})^{-1}(\mathcal{R}^T\ S_b^{2\mathcal{L}}\mathcal{R})\right)$$

where,

$$S_b^{2\mathcal{L}} = \sum_{i=1}^{C} t_i(\mu_i - \mu_o)^T\mathcal{L}\mathcal{L}^T(\mu_i - \mu_o)$$

and

$$S_w^{2\mathcal{L}} = \sum_{p=1}^{C}\sum_{q=1}^{t_i}\left(X_p^q - \mu_o\right)^T\mathcal{L}\mathcal{L}^T\left(X_p^q - \mu_o\right)$$

The optimal $\mathcal{L}$ and $\mathcal{R}$ values are computed by solving the generalized eigenvalue problems, $S_w^{2\mathcal{R}}x = \lambda S_b^{2\mathcal{R}}x$ and $S_w^{2\mathcal{L}}x = \lambda S_b^{2\mathcal{L}}x$ respectively. Since, $S_w^{2\mathcal{R}}$ and $S_w^{2\mathcal{L}}$ are nonsingular, $\mathcal{L}$ and $\mathcal{R}$ can be computed by computing an eigen-decomposition on $(S_w^{2\mathcal{R}})^{-1}S_b^{2\mathcal{R}}$ and $(S_w^{2\mathcal{L}})^{-1}S_b^{2\mathcal{L}}$. These eigen-decomposition works on rxr and cxc –dimension matrices, which are much smaller than LDA. Hence, 2D-LDA is more efficient than LDA in terms of memory occupation and time consumption.

III. KANNADA

Kannada is the official language of the Southern Indian state of Karnataka. It is a Dravidian language spoken by more than 50 million people in the Indian states of Karnataka, Andhra Pradesh, Tamil Nadu and Maharashtra.

A. Kannada Script

Kannada language has its own writing system, called Kannada script written from left to right and it is the descendents of Bramhi script. It is phonetic and uses fifty characters corresponding to 50 phonemes. The set of Kannada basic characters is known as Kannada Varnamale. Kannada varnamale consists of fifty characters having 16 vowels and 34 consonants. All consonants in Kannada have an inherent vowel and hence Kannada is an alphasyllabary.

The varnamale set along with their ITRANS (Indian language TRANSliteation) are shown in Figure 1. Though the vowel IÆ (RU) is not considered as part of the set during recent days, present work considers that so that the recognition engine is capable of recognizing older script also.

B. Data collection

Handwritten Kannada Varnamale data is collected using Tablet PC from native Kannada writers. Writers are advised to write in their own writing style without exercising any restrictions. A sample of handwritten Kannada varnamale is shown in Fig. 2.

C. Preprocessing

The collected raw data consists of the x- and y-coordinate values corresponding to the pen-tip movement. The raw data is subjected to noise removal and re-sampled in space along the arc length by linear interpolation so that each character has 30 equal number of points normalization. The re-sampled data are shifted and size normalized to fit in a square box of side length 1.

Vowels	ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ	
	a	aa	i	I	u	U	Ru	RU	
	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ	ಅಂ	ಅಃ	
	e	E	ai	o	O	ou	aM	aH	
Consonants	ಕ	ಖ	ಗ	ಘ	ಙ				
	ka	kha	ga	gha	nga				
	ಚ	ಛ	ಜ	ಝ	ಞ				
	ca	cha	ja	jha	nja				
	ಟ	ಠ	ಡ	ಢ	ಣ				
	Ta	Tha	da	dha	Na				
	ತ	ಥ	ದ	ಧ	ನ				
	ta	tha	da	dha	na				
	ಪ	ಫ	ಬ	ಭ	ಮ				
	pa	pha	ba	bha	ma				
	ಯ	ರ	ಲ	ವ	ಶ	ಷ	ಸ	ಹ	ಳ
	ya	ra	la	va	sha	Sha	sa	ha	La

Fig. 1. Kannada Character Set with their ITRANS shown below the character

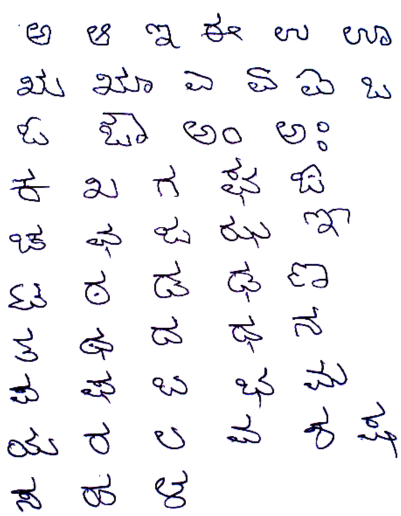

Fig. 2. Sample of Handwritten Kannada varnamale set

Let the pre-processed character data be represented by the sequence:

$$P = [p_1, p_2, \ldots\ldots\ldots \ldots p_{30}]$$

Where, the vectors $p_i = (a_i, b_i)^T$ are given by:

$$a_i = (x_i - x_{\min})/(x_{\max} - x_{\min})$$

$$b_i = (y_i - y_{\min})/(y_{\max} - y_{\min})$$

where, $(x_{\min}, y_{\min})^T$ and $(x_{\max}, y_{\max})^T$ denote the minimum and the maximum horizontal and vertical coordinate values of the character data under consideration.

IV. FEATURE EXTRACTION

To compare the performance of 2D-LDA with different feature combinations, following features are extracted from the size normalized sequence P.

A. Normalized x and y-Coordinates – F_1

The size normalized x and y-coordinate values given in a sequence P.

B. The trajectory feature - F_2

The trajectory feature consists of distance and angle features, corresponding to each of the size normalized points of a character. A normalized distance feature vector D and an angle feature vector θ, are computed with reference to the origin, from the sequence P.

Let

$$D = \{d_1, d_{2,} d_3, \ldots\ldots d_{30}\}$$

$$\theta = \{\theta_1, \theta_2, \theta_3, \ldots\ldots\ldots \theta_{30}\}$$

where,

$$d_i = \frac{\sqrt{a_i^2 + b_i^2}}{\max(\sqrt{a_i^2 + b_i^2})}$$

$$\theta_i = \frac{\arg(a_i + jb_i)}{\max(\arg(a_i + jb_i))}$$

C. First derivative feature - F_3

The first derivative features, $\hat{a}'$ and $\hat{b}'$ are calculated as given below:

$$a' = \frac{\sum_{i=1}^{2} i(a_{i+1} - a_{i-1})}{2\sum_{i=1}^{2} i^2}$$

$$b' = \frac{\sum_{i=1}^{2} i(b_{i+1} - b_{i-1})}{2\sum_{i=1}^{2} i^2}$$

$$\hat{a}' = \frac{a_i'}{\sqrt{a_i'^2 + b_i'^2}}$$

,

$$\hat{b}' = \frac{b_i'}{\sqrt{a_i'^2 + b_i'^2}}$$

D. Second derivative feature -F_4

The second derivative features are calculated from the expression of first derivative by replacing a and b by $\hat{a}'$ and $\hat{b}'$.

E. Deviation feature - F_5

The deviations of normalized sample points of a character from its centroid towards the x- and y-axes are calculated using the following equations and used as features. A vector consisting of deviation feature is used in training and testing of a second-stage classifier in a two-stage classifier to discriminate the confused numerals.

$$X_d = (a_{d1}, a_{d2}, a_{d3}, \ldots\ldots. a_{d30})$$
$$Y_d = (b_{d1}, b_{d2}, b_{d3}, \ldots\ldots\ldots .b_{d30})$$

where,

$$a_{di} = (a_i - \mu_a)/\sigma_b, \qquad b_{di} = (b_i - \mu_b)/\sigma_a$$

$$\mu_a = (1/30)\sum_{i=1}^{30} a_i, \qquad \mu_b = (1/30)\sum_{i=1}^{30} b_i$$

$$\sigma_a = \sqrt{(1/29)\sum_{i=1}^{30}(\mu_a - a_i)^2}$$

$$\sigma_b = \sqrt{(1/29)\sum_{i=1}^{30}(\mu_b - b_i)^2}$$

E. Estimate feature - F_6

The equations given below are used to calculate the feature values from sequence P and are called as the normalized estimate features.

$$est_{ai} = \frac{a_{ei} - a_{ei-min}}{a_{ei-\max} - a_{ei-\min}}$$

$$est_{bi} = \frac{b_{ei} - b_{ei-\min}}{b_{ei-\max} - b_{ei-\min}}$$

where,

$$a_{ei} = ((a_i - a_{i-1}) + (a_{i+1} - a_{i-1})/2)/2$$

$$b_{ei} = ((b_i - b_{i-1}) + (b_{i+1} - b_{i-1})/2)/2$$

$a_{ei-\min}$ and $a_{ei-\max}$ denote the minimum and maximum values in est_{ai}. Similarly, $b_{ei-\min}$ and $b_{ei-\max}$ denote the minimum and maximum values in est_{bi}.

V. EXPERIMENTS AND RESULTS

Writer independent experiments are carried out to compare the recognition performance of 2D-LDA for different feature combinations. The varnamale data collected from different users is divided into two sets, belonging to different, disjoint sets of writers. For training, 70% of the database is used and the remaining 30% is used for testing. Out of 5300 character samples, 3750 samples are used for training and 1550 samples are used for testing.

Though experiments are carried out up to four different feature combinations, feature combinations which yielded an average recognition accuracy of greater than or equal to 86.5% are only reported in this paper. The average recognition accuracies achieved for all the 50 classes with selected feature combinations are given in Table I. It is evident from Table I that, for second derivative and an estimate feature combination, maximum recognition accuracy of 88.3% is achieved.

Table I: Writer independent handwritten varnamale recognition performance with different feature combinations (Training samples: 3750; Test samples: 1550)

Sr. No.	Feature Combinations	Average Recognition Accuracy (%)
1	F_1 and F_3	87.2
2	F_1 and F_5	86.5
3	F_1 and F_6	87.3
4	F_1 and F_4	86.8
5	F_3 and F_6	88.1
6	F_3 and F_2	86.8
7	F_3 and F_4	85.9
8	F_4 and F_6	88.3
9	F_1 , F_3 and F_5	87.3
10	F_3 , F_4 and F_6	87.2

In addition to the above experiments, keeping second derivative and an estimate feature combination, three more experiments are carried out to achieve better average recognition accuracy. In the first step, the uniform sampling of each character in space is varied from 15 to 60 number of points. For 30 number of sampling points, best recognition accuracy of 88.3% is achieved. In the second step, for 30 sampling points and second derivative and estimate feature combination, percentage of training data size is varied to know the variation in an average recognition accuracy. By fixing the test data size of each class to 30% of the total collected data, the percentage of training data is varied from 10% to 100% in steps of 10%. The corresponding average recognition accuracies obtained from this experiment is given in Figure 3. It is evident from Figure 3 that, for the training data size of 60% to 100%, the average recognition accuracy is varying between 86% to 88.3%.

As per our knowledge, the proposed work is the first attempt made in recognizing online handwritten Kannada varnamle. So, we could not compare our result with any other research work.

VI. CONCLUSION

2D-LDA based writer independent recognition of online handwritten Kannada varnamale has been carried out. Experiments are carried out with different feature combinations to evaluate the performance of 2D-LDA. Second derivative feature and an estimate feature combination yielded a best average recognition accuracy of 88.3%. Experiments are also carried out for varying percentage of training data and varying space sampling of characters. As part of our future work, experimentation of the writer dependent case and adaptability of the recognition system for any single writer will be investigated.

ACKNOWLEDGEMENTS

The authors thank Technology Development for Indian Languages (TDIL) Program, Department of Information Technology, government of India for funding the collection of the Kannada online handwritten data used in this work. The authors wish to express their sincere thanks to Prof. A.G. Ramakrishnan of Indian Institute of Science, Bangalore for his valuable guidance and also for giving an opportunity to the first author to work in his MILE laboratory. Thanks are also due to UGC for giving Teacher Fellowship to the first author of the paper to carry out research work.

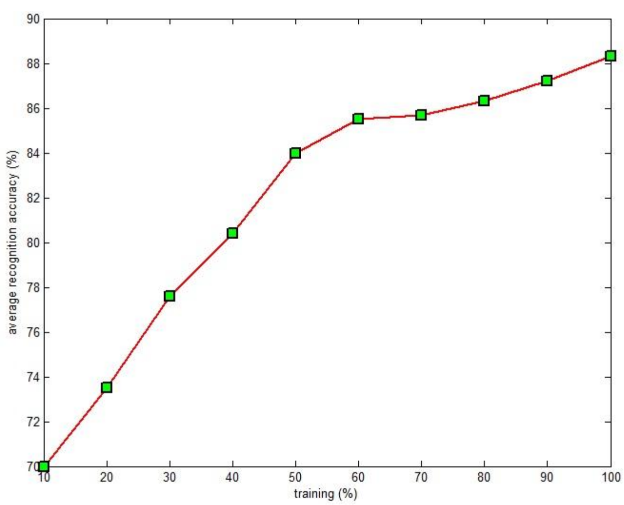

Fig. 3. Average recognition accuracy for varying training data size and fixed test sample size (Training data size = 70% and test data size = 30%)

REFERENCES

[1] Turk, M., Pentland, A., *"Eigenfaces for recognition"*, J. Cognitive Neurosci., 1991, 3 (1), pp. 71–86.

[2] Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J., *"Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection"*,. IEEE Trans. Pattern Anal. Machine Intell., 1997, 19 (7), pp. 711–720.

[3] R.O. Duda, P.E. Hart, and D. Stork, "*Pattern Classification*", Second Edition, John Wiley and Sons (Asia) Pvt. Ltd., 2006.

[4] Fengxi Song, Shuhai Liu, Jingyu Yang, "*Orthogonilzed Fisher disriminant*", Pattern Recognition, 2005, Vol. 38, pp. 311-313.

[5] FRIEDMAN, J., *"Regularized discriminant analysis"* Journal of the American Statistical Association, 1989, Vol. 84, pp.165–75.

[6] Yang, J., Zhang, D., Frangi, A.F., Yang, J.-y., "Twodimensional PCA: A new approach to appearance-based face representation and recognition", IEEE Trans. Pattern Anal. Machine Intell. 2004, Vol. 26 (1), pp. 131–137.

[7] Li, M., Yuan, B., "2D-LDA: A novel statistical linear discriminant analysis for image matrix". Pattern Recognition Lett., 2005, Vo. 26 (55), pp. 527–532.

[8] Ye, Jieping, Janardan, R., Park, C.H., Park, H., *"An optimization criterion for generalized discriminant analysis on undersampled problems"*. IEEE Trans. Pattern Anal. Machine Intell., 2004, Vol. 8, pp. 982–994.

M. Mahadeva Prasad
He did his master's degree in Electronics from Bangalore University in 1997. From 1997 to 1999, he worked as a faculty in the Department of Electronic Science, Bangalore University, Bangalore, India. In 1999, he joined Department of Studies in Electronics, Post Graduate Centre, Hemagangothri, University of Mysore, Hassan, India. Currently, he is working as an Assistant Professor and also the Course Coordinator of the Department. He is working as a member in BOE and BOS committees in Electronics of number of Universities. He has attended many national and international conferences and has published two papers in the field of online Handwritten Kannada characters in two international conferences. Now he is pursuing research in the field of online handwritten character recognition towards his doctoral degree. His research interests are, pattern recognition, character recognition, signal processing and communication Engineering.

M. Sukumar
He did his BE in E&C from University of Mysore in 1975, ME in Applied Electronics from MIT, Madras in 1979 and Ph.D. in Bio-Medical Engineering from IIT, Madras Currently, he is working as Professor and Head of the Department of Instrumentation Technology, S. J. College of Engineering, Mysore, Karnataka, India. He has 4 Journal papers and 12 conference papers to his credit and has attended many national and international conferences. He successfully guided two Ph.D. candidates and two more students are working with him. He successfully completed one major research project in the field of Bio-medical Engineering granted by LSRB, New Delhi, India. He served and working as member of various academic boards like, BOE and BOS. His research interests are, Bio-medical Engineering, MEMS technology, pattern recognition and signal processing.