Incremental Subclass Discriminant Analysis: A Case study in Face Recognition

Hemank Lamba, Tejas I. Dhamecha, Mayank Vatsa and Richa Singh
IIIT-Delhi

Abstract

Subclass discriminant analysis is found to be applicable under various scenarios. However, it is computationally very expensive to update the between-class and within-class scatter matrices. This research presents an incremental subclass discriminant analysis algorithm to update SDA in incremental manner with increasing number of samples per class. The effectiveness of the proposed algorithm is demonstrated using face recognition in terms of identification accuracy and training time. Experiments are performed on the AR face database and compared with other subspace based incremental and batch learning algorithms. The results illustrate that Incremental SDA yields significant reduction in time compared to SDA along with improving the accuracy compared to other incremental approaches.

Index Terms

Incremental, SDA, Subclass, Face Recognition, Face Identification

I. Introduction

Linear discriminant analysis (LDA) models the variability in intra-class and inter-class distributions to improve the classification performance. However, it assumes that underlying data follows normal distribution which may not not always be the case. As an extension to LDA, Zhu and Martinez [1] proposed subclass discriminant analysis (SDA). They showed that when the underlying data from the same class conforms to multiple normal distributions, it is useful to consider each of them as a subclass. The classification time of SDA is linearly proportional to the number of subclasses and the number of features. The property of reduced time complexity for classification makes it more applicable to real time scenarios. In the original formulation of LDA and SDA, if a new gallery image is to be added, it is necessary to recompute the between-class and within-class scatter matrices. This results in a monolithic architecture and makes it computationally very expensive to update the discriminant vectors using only the *new* samples being added and hence it is not *incremental* in nature. Research has been done to formulate incremental LDA (ILDA) [2], [3], [4]. However, to the best of our knowledge no research has been done to formulate incremental SDA (ISDA). Main contribution of this research is developing an incremental formulation of SDA to incorporate the information obtained from updated samples per class. The effectiveness of the proposed algorithm is evaluated for face recognition application and *identification accuracy* as well as *training time* are used as the performance metrics on AR face database. Section 2 describes the formulation of SDA followed by the proposed approach for incremental SDA in Section 3. Section 4 presents the experiments performed, results achieved and the analysis. Section 5 includes the conclusion and future directions.

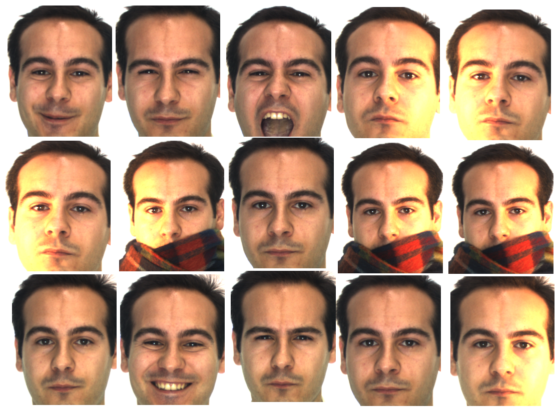


Fig. 1. Sample images from the AR face database illustrating the variation in face images of the same person [5].

II. Subclass Discriminant Analysis

Discriminant analysis (DA) techniques have a fundamental criterion called the Fisher-Rao's criterion [6]

$$J(\mathbf{v}) = \frac{|\mathbf{v^T A v}|}{|\mathbf{v^T B v}|} \tag{1}$$

where $\mathbf{A}$ represents the between-class variability and $\mathbf{B}$ represents the within-class variability. The goal is to find the projection direction $\mathbf{v}$ which differentiates the classes optimally. The projection direction which leads to minimum possible within-class variance and maximum possible between-class variance is the most discriminative direction $\mathbf{v}_{opt}$.

$$\mathbf{v}_{opt} = \underset{\mathbf{v}}{\operatorname{argmin}}\, J(\mathbf{v}) = \underset{\mathbf{v}}{\operatorname{argmin}} \frac{|\mathbf{v^T S}_B \mathbf{v}|}{|\mathbf{v^T S}_W \mathbf{v}|} \tag{2}$$

Different DA techniques modify the definitions of $\mathbf{A}$ and/or $\mathbf{B}$. For example, LDA uses between-class and within-class scatter matrix as $\mathbf{A}$ and $\mathbf{B}$, respectively. SDA defines $\mathbf{A}$ as

$$\mathbf{S}_B = \sum_{i=1}^{c-1} \sum_{j=1}^{H_i} \sum_{k=i+1}^{c} \sum_{l=1}^{H_k} p_{ij} p_{kl} (\mu_{ij} - \mu_{kl})(\mu_{ij} - \mu_{kl})^T \tag{3}$$

where c is the number of classes and H_i is the number of subclasses in the i^{th} class. μ_{ij} is the mean of the j^{th} subclass of i^{th} class, and p_{ij} is the prior probability of j^{th} subclass of i^{th} class. The techniques to divide a class into subclasses and to find the value of H_i are discussed by Zhu and Martinez in [1]. The matrix $\mathbf{B}$ is formulated as the within-class scatter matrix and defined as

$$\mathbf{S}_W = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T \tag{4}$$

where n_i is the number of samples in i^{th} class and $n = \sum_{i=1}^{c} n_i$. On a different note, according to Fukunaga [7] it should be taken into account that in Eq. 2 $J(\cdot)$ can take the form of

$$J(\mathbf{v}) = \frac{|\mathbf{v^T S}_B \mathbf{v}|}{|\mathbf{v^T}(\mathbf{S}_B + \mathbf{S}_W)\mathbf{v}|} = \frac{|\mathbf{v^T S}_B \mathbf{v}|}{|\mathbf{v^T S}_T \mathbf{v}|} \tag{5}$$

without loosing generality.

III. Incremental Subclass Discriminant Analysis

Over the years, many different approaches have been proposed for incremental LDA [2], [3] and they vary in terms of approaches to update the within class and between class scatter matrices. In this research, a sufficient spanning sets [2] approach is followed to develop an ISDA algorithm. Sufficient spanning set is a set of basis vectors that span the space of most data variations.

Let d_1 be the number of data samples (contained in initial training set) from which the first covariance matrix $\mathbf{C}_1$ is created, and d_2 be the number of data samples (contained in incremental training set) from which the second covariance matrix $\mathbf{C}_2$ is created. If the *new* covariance matrix $\mathbf{C}_m$ of the merged dataset (initial training set + incremental training set) is obtained using all the $d_1 + d_2$ samples, then the time complexity of computing $\mathbf{C}_m$ turns out to be $O(N^2(d_1 + d_2))$, where N is the number of features. Hall et al. [8] proposed a way to *merge* two covariance matrices as an alternative to compute a *new* covariance matrix. It focuses on finding an eigenspace spanned by (possibly less number of) basis vectors in which the representation of the merged dataset is possible with sufficiently good approximation. This set of basis vectors of eigenspace (eigenvectors) is called the sufficient spanning set. If $\mathbf{C}_i$ is represented using eigenmodels $\{\mu_i, N_i, Ev_i, \Lambda_i\}$ $(i \in \{1, 2\})$, where μ_i is the mean of the i^{th} dataset, N_i is the number of samples in the i^{th} dataset, Ev_i is the set of (selected first few) eigenvectors of $\mathbf{C}_i$, and Λ_i is the matrix containing eigenvalues corresponding to the eigenvectors in Ev_i, then the merged eigenmodel obtained using sufficient spanning set will be $\{\mu_m, N_m, Ev_m, \Lambda_m\}$ which can be obtained using Eqs. 6 to 9.

$$N_m = (N_1 + N_2) \tag{6}$$

$$\mu_m = (N_1\mu_1 + N_2\mu_2)/(N_m) \tag{7}$$

$$\mathbf{C}_m = \tfrac{N_1}{N_m}\mathbf{C}_1 + \tfrac{N_2}{N_m}\mathbf{C}_2 + \tfrac{N_1 N_2}{N_m^2}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

$$\mathbf{\Phi} = QRDecomposition([Ev_1, Ev_2, (\mu_1 - \mu_2)])$$

$$R = EigenVectors(\mathbf{\Phi}^T \mathbf{C}_m \mathbf{\Phi})$$

$$\Lambda_m = EigenValues(\mathbf{\Phi}^T \mathbf{C}_m \mathbf{\Phi}) \tag{8}$$

$$Ev_m = \mathbf{\Phi} R \tag{9}$$

It is interesting to note here that Φ is the sufficient spanning set of eigenvectors which are the basis of eigen decomposition of the modified covariance matrix C_m. Using this technique of merging two covariance matrices, Kim et al. [2] formulated the extension to LDA. That is, having found eigenmodels of between-class scatter matrices, merge them using the sufficient spanning set of matrix $[\mathbf{S}_{B,1}, \mathbf{S}_{B,2}, \mu_2 - \mu_1]$, where $\mathbf{S}_{B,i}$ $(i \in \{1, 2\})$ is the between-class scatter matrix of the i^{th} dataset. In the similar manner, merged total scatter matrix is found using the sufficient spanning set of matrix $[\mathbf{S}_{T,1}, \mathbf{S}_{T,2}, \mu_2 - \mu_1]$, where $\mathbf{S}_{T,i}$ $(i \in \{1, 2\})$ is the within-class scatter matrix of the i^{th} dataset. Here the definition of $\mathbf{S}_{B,i}$ is as follows.

$$\mathbf{S}_{B,i} = \sum_{k=1}^{C} n_k (m_k - \mu)(m_k - \mu)^T \tag{10}$$

where n_k is the number of samples in k^{th} class, m_k is the mean of samples belonging to k^{th} class, μ is the mean of data samples, and C is number of classes.

To formulate incremental SDA algorithm, we propose to compute $\mathbf{S}_{B,i}$ using Eq. 3 which defines between-class scatter matrix of the i^{th} dataset and helps in incorporating inter-subclass variations. Since the subclass labels are not available for the incremental batch, it becomes challenging to compute $\mathbf{S}_{B,i}$ as defined in Eq. 3. Ground truths of subclass labels are unknown, therefore it is necessary to use an unsupervised clustering technique to find subclass labels of new samples (the samples from incremental training set). In this research, we propose to use nearest neighbour (NN) [9] clustering technique to find the subclass labels of the sample. Once the subclass labels are assigned to the samples, they can now be used to compute the scatter matrix $\mathbf{S}_{B,2}$. Similarly, total scatter matrix $\mathbf{S}_{W,2}$ for the new batch of training samples can also be computed. Using the mathematical formulation explained in Eq. 6, 7, 8, and 9 eigenmodel $\{\mu_m, N_m, EV_{B,m}, \Delta_{B,m}, n_{m,j}, \alpha_{m,j} | j = 1, 2, \ldots, c\}$ of incremented between-subclass scatter matrix $\mathbf{S}_{B,m}$ and the eigenmodel $\{\mu_m, N_m, EV_{T,m}, \Lambda_{T,m}\}$ of incremented total scatter matrix $\mathbf{S}_{T,m}$ can be found. ($\Delta_{B,m}$ and $\Lambda_{T,m}$ are the matrices containing eigenvalues of the corresponding eigenvectors. $n_{m,j}$ and $\alpha_{m,j}$ are the number of samples and the matrix containing coefficients of j^{th} class.) The procedure for finding the discriminative components U from the given eigenmodels of $\mathbf{S}_{B,m}$ and $\mathbf{S}_{T,m}$ is as follows [2].

$$
\begin{aligned}
U &= Z\Omega R_D && \text{where,} \qquad (11)\\
Z &= \mathbf{S}_{T,m}\Lambda_{T,m}^{-\frac{1}{2}}, \\
\Omega &= QRDecomposition(Z^T\mathbf{S}_{B,m}), && \text{and} \\
R_D &= EigenVectors(\Omega^T Z^T \mathbf{S}_{B,m}\Delta_{B,m}\mathbf{S}_{B,m}^T Z\Omega)
\end{aligned}
$$

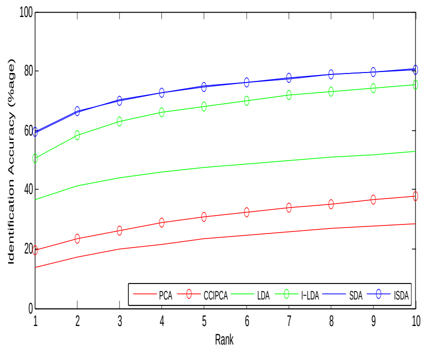
IV. Experiments and Results

To evaluate the performance of the proposed approach, the experiments are performed on the AR face database [5] and results are compared with LDA [10], ILDA [2], PCA [10], CCIPCA [11], and SDA [1]. The database consists of more than 4,000 color face images of 126 subjects. In the context of ISDA, we will consider subjects as classes amongst which the classification is to be done. The database consists of frontal face images with challenges like illumination, expression, and occlusion (scarves and glasses). In the experiments, we have used images of 119 classes and 26 images per class, which resulted in the overall database of the size 3094.

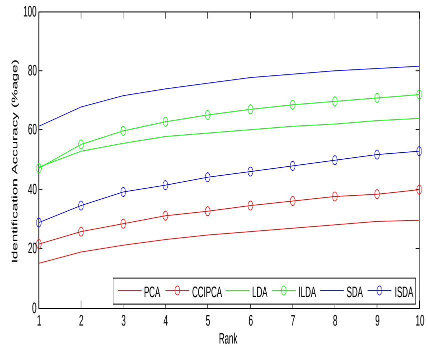
Faces are detected using the AdaBoost face detector available in OpenCV and resized to 29 $\times$ 21 pixels. These pixel intensity values are used as the features for classification. The database is divided into 50% training and 50% testing. 13 (randomly selected) images of each individual are selected for training and the remaining 13 are used for testing. To evaluate the performance of incremental learning, the training set is further divided into three splits. Batch I consists of 1071 images (9 images per subject $\times$ 119 subjects) whereas batch II and batch III contain 238 images (2 images per subject $\times$ 119 subjects) each. The incremental approach is initially trained with batch I and the performance is evaluated using with the whole testing set. In the next step, incremental training is performed with batch II and batch III successively. For each incremental training, the performance is evaluated on the overall testing set. It should be noted that in this case study, no new classes are being included during incremental training; only the number of images per class is being updated. To compare the performance, non-incremental algorithms are also evaluated by training with

- batch I only
- combined batch I and batch II, and
- combined batch I, batch II and batch III.

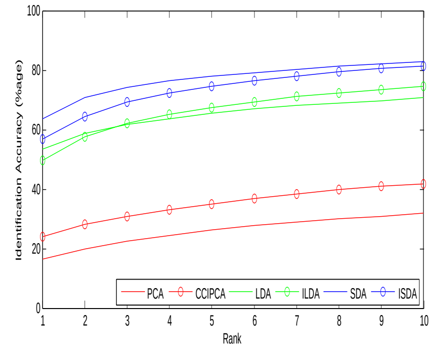
The performance of both offline and online learning algorithms is evaluated on the overall testing set in terms of both training time and identification accuracy. The results reported in Fig. 2 are achieved using the protocol described above and with five times random cross validation. PCA experiments are performed with 100 principal components.



(a) Batch I (Initial)



(b) Batch II (Update)



(c) Batch III (Convergence)

Fig. 2. CMC plots of the proposed ISDA and performance comparison with PCA, CCIPCA, LDA, ILDA, and SDA. The results are computed for (a) training with batch I, b) incremental training with batch II, and c) incremental training with batch III.

- **Analysis**: The results in Fig. 2 and Table I show that the proposed ISDA outperforms existing approaches. The error bars in Fig. 3 show that the standard deviation across different trials is also very small. With initial batch training, no incremental update has been performed, therefore ISDA behaves exactly same as SDA.

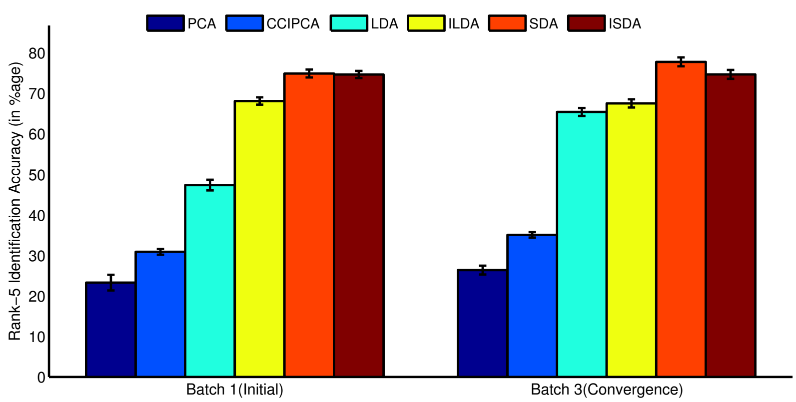


Fig. 3. Rank 5 accuracy of incremental and non-incremental algorithms with five cross validation trials. The error bars show the standard deviation for every algorithm.

- **Time**: Table I shows that the overall turn-around-time (training time + testing time) of ISDA in each batch increment is significantly less than that of SDA.

	PCA	CCIPCA	LDA	ILDA	SDA	ISDA
Initial batch (batch I)	69.6	18	15.2	11.4	**6167.7**	**6980.7**
Batch II	83.3	22.3	19.5	13.5	**10610**	**28.4**
Batch III	99.8	26.3	20.2	25.4	**17494**	**30.9**

TABLE I

INCREMENTAL TIME TAKEN (IN SECONDS) BY EACH OF THE APPROACHES FOR INITIAL TRAINING (BATCH I) AND INCREMENTAL TRAININGS (BATCH II AND BATCH III)

- **Relevance**: Table II describes the co-occurrence of correct classifications (✓) and/or misclassifications (✗) between SDA and ISDA. It turns out that for only $\frac{110+60}{1547} \times 100\% = 10.98\%$ of the times, the decisions taken by SDA and ISDA differ. (Confusion matrix is shown for batch III at rank 5.)

Confusion matrix @ Rank 5		**SDA** ✓	**SDA** ✗
ISDA	✓	1068	60
ISDA	✗	110	309

TABLE II

CONFUSION MATRIX FOR COMPARING THE PERFORMANCE OF SDA AND ISDA. ✓ AND ✗ REPRESENT THE CORRECTLY CLASSIFIED AND MISCLASSIFIED SAMPLES RESPECTIVELY. THE NUMBERS IN EVERY CELL REPRESENT THE CO-OCCURRENCE OF DECISIONS (CORRECT/WRONG) TAKEN BY SDA AND ISDA. FOR EXAMPLE, ✓✓ BLOCK SHOWS THAT FOR 1068 SUBJECTS, BOTH SDA AND ISDA GAVE CORRECT DECISIONS AT RANK 5.

- **Iterative incremental training**: One interesting observation is that with the first incremental batch, the accuracy of ISDA decreases while with successive incremental training the accuracy is at par with SDA. It is our assertion that incremental training with smaller data may be introducing noisy basis vectors, which get improvised with successive training. However more investigation is required to understand this result.

V. CONCLUSION

This research presents incremental subclass discriminant analysis approach using sufficient spanning sets. The proposed ISDA algorithm is evaluated with the application of face recognition. The results on the AR face database suggest that incremental SDA is over two times faster than SDA with almost similar rank-5 identification accuracy. Though the initial results are very promising, verifying the generalizability of ISDA to other pattern classification problems still needs to be explored.

(a) (b)

Fig. 4. Face images (a) correctly classified by SDA but misclassified by ISDA and (b) correctly classified by ISDA but misclassified by SDA.

ACKNOWLEDGEMENT

The authors would like to thank Dr. A. Martinez for providing access to the AR face database. This research is supported through a grant from Department of Information Technology, Government of India.

REFERENCES

[1] M. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1274 –1286, 2006.

[2] T. K. Kim, S. F. Wong, B. Stenger, J. Kittler, and R. Cipolla, "Incremental linear discriminant analysis using sufficient spanning set approximations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[3] J. T. Kwok and H. Zhao, "Incremental eigen decomposition," in *International Conference on Artificial Neural Networks*, 2003, pp. 270–273.

[4] H. Zhao and P. C. Yuen, "Incremental linear discriminant analysis for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 38, no. 1, pp. 210–221, 2008.

[5] A.M. Martinez and R. Benavente, "The AR face database," *CVC Technical Report #24*, 1998.

[6] R. A. Fisher, "The use Of multiple measurements In taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[7] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*, Academic Press, 1990.

[8] P. Hall, D. Marshall, and R. Martin, "Merging and splitting eigenspace models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 1042–1049, 2000.

[9] L. Devroye, L. Gyorfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, Springer, 1996.

[10] 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.

[11] J. Weng, Y. Zhang, and W. Hwang, "Candid covariance-free incremental principal component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1034–1040, 2003.