INCREMENTAL SUBCLASS DISCRIMINANT ANALYSIS: A CASE STUDY IN FACE RECOGNITION

Hemank Lamba, Tejas Indulal Dhamecha, Mayank Vatsa and Richa Singh

IIIT-Delhi, India

ABSTRACT

Subclass discriminant analysis is found to be applicable under various scenarios. However, it is computationally expensive to update the between-class and within-class scatter matrices in batch mode. 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, compared to SDA, incremental SDA yields significant reduction in time along with comparable accuracy.

Index Terms— Incremental Learning, SDA, Subclass, Face Recognition

1. 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 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 which helps in improving the classification accuracy. Further, in SDA, it is observed that the classification time is linearly proportional to the number of subclasses and the number of features. Therefore, the property of reduced time complexity for classification makes it more applicable to real time scenarios. However, SDA, similar to other discriminant functions, is trained in batch mode, which may be time consuming. In other words, 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 expensive to update the discriminant vectors using only the *new* samples being added. Research has been done to formulate incremental LDA (ILDA) [2], [3], [4]. However, to the best of our knowledge no formulation exists for incremental SDA (ISDA).

The 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. 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.

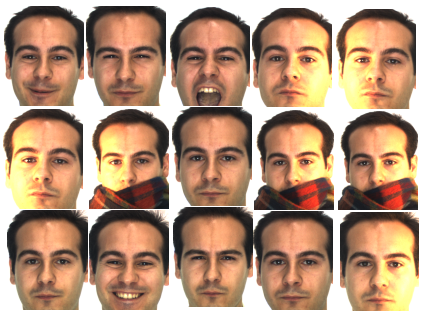


Fig. 1. Sample images from the AR face database [5] illustrating the variation in face images of the same person.

2. SUBCLASS DISCRIMINANT ANALYSIS

Discriminant analysis (DA) techniques follow a fundamental criterion called 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 between 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. On the other hand, 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 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

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

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, without loss of generality, $J(\cdot)$ can take the form of

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

3. INCREMENTAL SUBCLASS DISCRIMINANT ANALYSIS

It is important to understand how incremental training is different than including new samples in the gallery (i.e. *template update*). It is not difficult to visualize the cases where initial training may not cover *all* possible variabilities in the data. Therefore, the additional samples may change the intraclass and interclass variabilities. Due to this, it might not be sufficient to simply include these additional samples in the gallery and increase the number of *seen* samples without any *learning*. It is required that the underlying classifier *learns* these added variabilities, which can be incorporated only when classifiers decision boundary is modified using the newly added samples.

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 case of SDA, between-sub-class scatter matrix $\mathbf{S}_B$ and within-sub-class scatter matrix $\mathbf{S}_W$ have to be recomputed in incremental manner, to *learn* new samples. In this research, the sufficient spanning sets [2] approach is followed to develop the ISDA formulation. 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 computing *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 found using Eqs. 6 to 9.

$$N_m = (N_1 + N_2) \quad (6)$$

$$\mu_m = (N_1\mu_1 + N_2\mu_2)/(N_m) \quad (7)$$

$$\begin{aligned}\mathbf{C}_m &= \frac{N_1}{N_m}\mathbf{C}_1 + \frac{N_2}{N_m}\mathbf{C}_2 \\ &+ \frac{N_1N_2}{N_m^2}(\mu_1-\mu_2)(\mu_1-\mu_2)^T\end{aligned}$$

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

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

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

$$Ev_m = \mathbf{\Phi}R \quad (9)$$

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 $\mathbf{C}_m$. Using this technique of merging two covariance matrices, Kim et al. [2] formulated incremental LDA i.e. after calculating 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 calculated 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 total scatter matrix of the i^{th} dataset. Here the definition of $\mathbf{S}_{B,i}$ is as follows.

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

where n_k is the number of samples in the 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 define between-class scatter matrix $\mathbf{S}_{B,i}$ of the i^{th} dataset in the same way as it is defined by SDA in Eq. 3. This helps in incorporating inter-subclass variations. However, the subclass labels are not available for the incremental batch thus making it challenging to compute the between-subclass scatter matrix. We use an unsupervised clustering technique to find the sub-class of the new sample (a sample from the incremental training set). Here, the unsupervised clustering is more pertinent due to the fact that ground truth of sub-class labels are not known, only the ground truths of class labels are known.

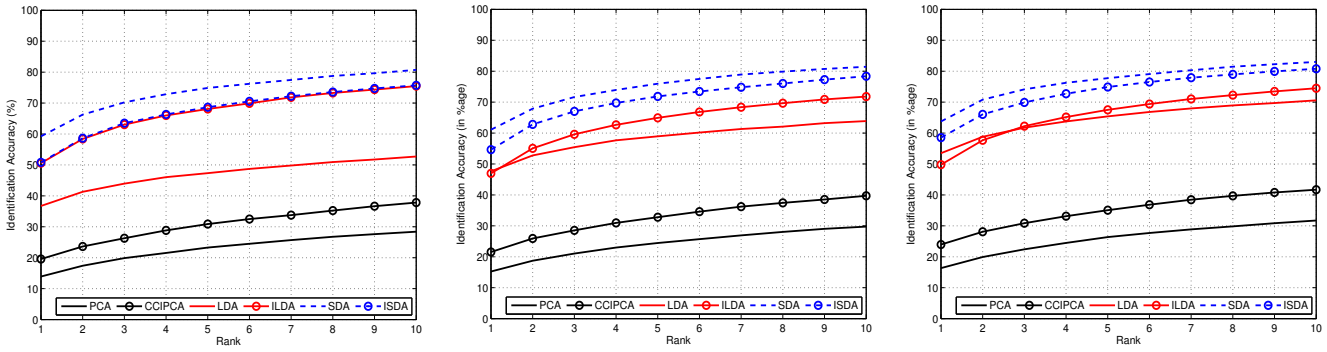
In this research, we propose to use nearest neighbor (NN) [9] clustering technique to find the sub-class labels of the sample. Once the sub-class labels are assigned to the samples, they can now be used to compute the scatter matrix $\mathbf{S}_{B,2}$. Similarly, total scatter matrix $\mathbf{S}_{T,2}$ for the new batch of training samples can also be computed. Using the mathematical formulation explained in Eqs. 6-9 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 calculated. $\Delta_{B,m}$ and $\Lambda_{T,m}$ are the matrices containing eigenvalues of the corresponding eigenvectors. $n_{m,j}$ and $\alpha_{m,j}$ are 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 followed [2].

$$U = Z\Omega R_D \quad \text{where} \quad (11)$$

$$Z = \mathbf{S}_{T,m}\Lambda_{T,m}^{-\frac{1}{2}} \quad \text{and}$$

$$\Omega = QRDecomposition(Z^T\mathbf{S}_{B,m}) \quad \text{and}$$

$$R_D = EigenVectors(\Omega^T Z^T \mathbf{S}_{B,m}\Delta_{B,m}\mathbf{S}_{B,m}^T Z\Omega) \quad (12)$$



(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, SDA. The results are computed for a) training with batch I (initial set), b) incremental training with batch II, and c) incremental training with batch III. Note that in batch I, performance of SDA is computed using [1] and performance of ISDA is computed using the proposed approach.

	PCA	CCIPCA	LDA	ILDA	SDA	ISDA
Initial batch (batch I)	69.6	18	15.2	11.4	**6167.7**	**6712**
Batch II	83.3	22.3	19.5	13.5	**10610**	**22.2**
Batch III	99.8	26.3	20.2	25.4	**17494**	**24.6**

Table 1. Incremental time taken (in seconds) by each of the approaches, for initial training (Batch I) and the incremental trainings (Batch II and III)

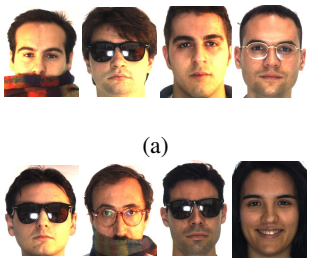
4. EXPERIMENTS AND RESULTS

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

Faces are detected using the AdaBoost face detector available in OpenCV, converted to grayscale, and resized to 29 $\times$ 21 pixels. These grayscale 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 while 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 tested 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 are being updated. To compare the performance, non-incremental algorithms i.e. PCA, LDA and SDA are also evaluated by training with

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

The performance of all the algorithms is compared 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 with five times random cross validation. Also, PCA experiments are performed with top 100 principle components. Key analysis and observations are as follows.



(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.

[1] Out of 126 subjects, face detection algorithm failed to detect few faces for seven subject and therefore, the experiments were performed with 119 subjects only.

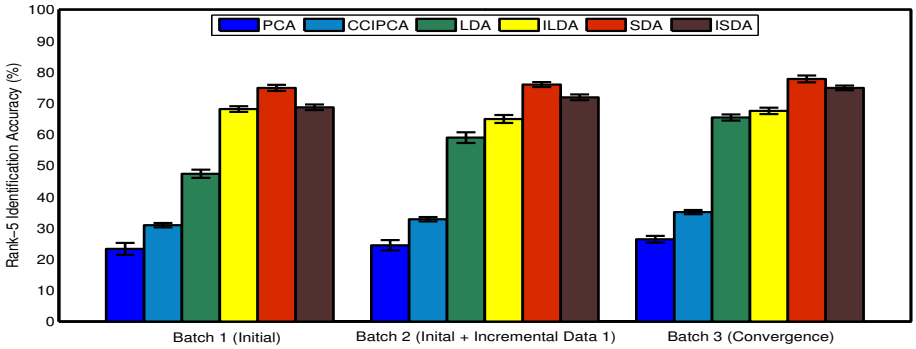


Fig. 3. Rank 5 accuracy of incremental and non-incremental algorithms with five cross validation trials. The error bars show the standard deviation for every algorithm.

- **Accuracy**: The results in Fig. 2 show that the proposed ISDA achieves identification accuracy which is comparable to that of SDA. The error bars in Fig. 3 show that the standard deviation across different trials is also very small.
- **Time**: Table 1 shows that the overall turn-around-time (training time + testing time) of ISDA in each batch increment is significantly less than that of SDA. It can be viewed as though a little amount of sacrifice in identification accuracy is made in order to achieve a huge gain in time complexity.
- **Relevance**: Table 2 describes the co-occurrence of correct classifications (✓) and/or misclassifications (✗) between SDA and ISDA. It turns out that for only $\frac{87+50}{1547} \times 100\% = 8.85\%$ of the times, the decisions taken by SDA and ISDA differ. Table 2 shows the confusion matrix for batch III at rank 5.

Confusion matrix @ Rank 5		SDA	
		✓	✗
ISDA	✓	1091	50
	✗	87	319

Table 2. 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 1091 samples, both SDA and ISDA gave correct decisions at rank 5.

- **Batchwise incremental training**: It is interesting to observe that with the successive batches of incremental training, difference between accuracy of ISDA and SDA is reduced. This suggests that with more batches of incremental training, the behavior of ISDA is closer to SDA. Fig. 2 and Table 2, collectively, also points the possibility that the projection vectors calculated using the proposed incremental approach tend to converge to projection vectors achieved in SDA.

5. CONCLUSION

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

6. ACKNOWLEDGEMENT

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

7. REFERENCES

[1] M. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE TPAMI*, 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 CVPR*, 2007, pp. 1–8.

[3] J. T. Kwok and H. Zhao, "Incremental eigen decomposition," in *ICANN*, 2003, pp. 270–273.

[4] H. Zhao and P. C. Yuen, "Incremental linear discriminant analysis for face recognition," *IEEE TSMC-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 TPAMI*, 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 TPAMI*, vol. 19, no. 7, pp. 711–720, 1997.

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