CHAPTER 1
INTRODUCTION

1.1 BASIC CONCEPTS OF FACE RECOGNITION

Database security has become a concern in many of the Government agencies. As a reliable security mechanism biometric based security systems are used [1]. Biometric system can be integrated into any application requiring security or access control. It can effectively eliminate risks associated with less advanced technologies that are based on what a person has or knows rather than who a person really is. Perhaps the most common biometrics are finger prints, Iris, voice, palm geometry and face. However, biometric recognition methods have drawbacks too. Iris recognition method is extremely accurate, but expensive to implement and it is not accepted by many people because of the difficulties involved. Finger print identification method is reliable and non-intrusive, but it is not suitable for non-collaborative individuals.

<table>
<thead>
<tr>
<th>Areas</th>
<th>Specific Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entertainment</td>
<td>Video games, Human-robot interaction, Human-computer interaction</td>
</tr>
<tr>
<td>Smart cards</td>
<td>Driver’s license, Entitlement programs, Immigration, National ID, Passports, Voter registration.</td>
</tr>
<tr>
<td>Information Security</td>
<td>Personal device logon, Application security, Database security, File encryption, Internet access.</td>
</tr>
<tr>
<td>Law enforcement and Surveillance</td>
<td>Advanced video surveillance, Portal control, Post event analysis, Suspect tracking and Investigation.</td>
</tr>
</tbody>
</table>
On the contrary, Face recognition method seems to be a good compromise between reliability and social acceptance, and also preserves security and privacy. It has a wide range of applications which are described in [2]. Some of its applications are listed in Table 1.1.

Face Recognition can be defined as the identification of individuals from the real time images of their faces captured when the individual is present and comparing the same with a stored database of faces labeled with individual’s identities. Generally, the Face recognition method can be classified as follows [3]:

(i) **Holistic method:** In this method, Statistical Analysis is performed to recognize a given face image and each face image is represented as a high dimensional vector by concatenating the gray levels of all pixels in the face image. The main advantage of this representation is that it implicitly preserves all the detailed information which is useful for distinguishing face images.

(ii) **Structural method:** In this method, local features of face image such as eyes, nose and mouth are extracted, and then their locations and local statistics are fed into a classifier. Moreover, a face image is represented by a set of its local features which reduce the dimension of the sample space. Since this method uses these limited features for representing a face image, the performance of recognition may be reduced.

(iii) **Hybrid method:** This method uses both structural and holistic features to recognize a face image which increases the recognition performance. However, this method is not popular since the holistic and structural features are individually sensitive to different variation factors.

Face recognition is a difficult problem as there are numerous factors such as 3D pose, facial expression, hair style, make up and etc., which affect the appearance of an individual’s facial features. In addition to these varying factors, the lighting, background, and scale changes make the recognition task even more challenging. In
addition to that, problematic conditions continue due to inferences of noise, occlusion, and many other possible factors [4]. Thus, there is a necessity to design an effective and reliable method for face recognition which will not be affected by the above mentioned factors. The next section presents the research contributions made so far for solving the problems of face recognition.

1.2 PRESENT SCENARIO

Research on human face recognition has been going on for a few decades and a number of methods have already been developed. The survey papers presented by Zhao et al. [2], Abate et al. [4] and Chellappa et al. [5] give an account of the published research work in this area. The methods developed during 1966-1989 focused on detecting individual features such as eyes, nose, mouth and head outline and defined a face model by the position, size and relationships between these features. Face recognition using the features detected had many problems in the accurate recognition of faces. This is due to the individual features and their relationships provided insufficient representation for face recognition [6].

Bledsoe was the first, to attempt semi automated face recognition process that classified faces on the basis of fiducial marks entered on photographs by hand [7,8]. The parameters used for the classification are the normalized distances and the ratios between the points such as eye corners, mouth corners, nose tip and chin point. Later, Goldstein et al. have proposed a Human face identification method which constructed a vector of up to 21 features, and recognized faces using standard Pattern classification techniques [9]. In this work, the researchers have chosen features like shade of hair, length of ears, and lip thickness as additional features for recognition.

Fischler et al. have attempted to measure similar features automatically [10]. They described a linear embedding algorithm which used local feature template matching and a global measure fit to find and measure facial features. This template matching approach has been continued and improved by Yuille et al. [11]. Their strategy was
based on "deformable templates", which are parameterized models of the face and its features in which the parameter values are determined by interactions with the image.

Automated face recognition method was proposed by Cannon S.R et al. Craw et al. and Wong et al. [12-14] by characterizing a face by a set of geometric parameters and performing face recognition based on these parameters. Kanade's face identification system was the first system in which all steps of the recognition process were automated [15]. This system calculated a set of facial parameters from a single face image and used a pattern classification technique to match the face from a known face database. This system is purely a statistical approach, depending primarily on local histogram analysis and absolute gray-scale values. Burt had proposed a smart sensing approach based on multi resolution template matching [16, 17]. This system works effectively under limited circumstances, but will suffer from the typical problems of correlation-based matching, including sensitivity to image size and noise. The above mentioned methods ignored the issue of what aspects of the face stimulus are important for identification, assuming that predefined measurements are relevant and adequate. To overcome this assumption, Turk et al. followed the information theory approach for coding and decoding of face images which emphasized the significant local and global features [18]. In this, the relevant information in a face image is extracted, encoded efficiently, and then compared with a database of models encoded already. This approach somehow captures the variation in a collection of face images and uses such information to encode and compare individual face images. This is a simple approach for extracting the information contained in a face image. Moreover Turk observed that the method for representation of face image of size w×h pixel as vectors in wh-dimensional space increases the computational complexity and this space is called as sample space whose dimension is very large [18]. Since face images have similar structure, the image vectors are correlated and any image in the sample space can be represented as a vector that contains only the variations in a collection
of face images. Thus, this vector can be represented in a lower-dimensional subspace without losing significant amount of information [18].

Turk and Pentland have proposed the Eigenface method for finding such a lower dimensional subspace [19]. The key idea behind the Eigenface method, which uses Principle Component Analysis (PCA) technique, is to find the best set of projection directions in the sample space \((W_{opt})_{PCA}\) that will maximize the total scatter across all images such that

\[
(W_{opt})_{PCA} = \arg \max_W |W^T S_r W|
\]  

(1.1)
is maximized, where \(S_r\) is the total scatter matrix of the training set samples and \(W\) is the matrix whose columns are the orthonormal projection vectors, and \(W^T\) is the transpose of \(W\) matrix. The projection directions are also called as eigenfaces. Any face image in the sample space can be approximated by a linear combination of the significant eigenfaces. The sum of the eigen values that correspond to the eigenfaces are not used in reconstruction which produced the mean square error of reconstruction. This method is an unsupervised technique since it does not consider the classes within the training set data. Additionally, the criterion does not attempt to minimize the within-class variation that leads to overlapping between the classes. Belhumeur et al. [20] and Swets et al. [21] have proposed the Linear Discriminant Analysis (LDA) method to overcome the limitations of the eigen face method by applying the Fisher’s Linear Discriminant (FLD) criterion. This criterion tries to maximize the ratio

\[
(W_{opt})_{FLD} = \arg \max_W \frac{|W^T S_b W|}{|W^T S_w W|}
\]  

(1.2)

where \(S_b\) is the between class scatter matrix, and \(S_w\) is the within class scatter matrix. Thus, by applying this method, the optimal projection \(W_{opt}\) is chosen as the matrix with orthonormal columns that maximizes the ratio of the determinant of the between class scatter matrix of the projected samples to the determinant of the within class scatter matrix of the projected samples. In face recognition process, this method cannot be applied directly since the dimension of the sample space is
typically larger than the number of samples in the training set. As a consequence, \( S_w \) is singular in this case. This problem was identified and named as **Small Sample Size (SSS)** problem by K. Fukunaga et al. [22].

In the last decade, numerous methods have been proposed to solve SSS problem. Tian et al. [23] used the pseudo inverse method by replacing \( S_w^{-1} \) with its pseudo inverse. The Perturbation method had been used in [24] where a small perturbation matrix \( \Delta \) is added to \( S_w \) in order to make it non singular. However, these methods are computationally expensive since the scatter matrices are very large. Swets and Weng [21] have proposed a two stage PCA+LDA method, which is known as Fisher Face method in which PCA is used for dimension reduction so as to make \( S_w \) nonsingular before the application of LDA. However, in order to make \( S_w \) nonsingular, some directions corresponding to the small Eigenvalues of \( S \), are thrown away in the PCA step. Thus, applying PCA for dimensionality reduction has the potential to remove dimensions that contain discriminative information [25-27]. Chen et al. [28] have proposed the Null space method based on the modified Fisher’s Linear Discriminant Criterion

\[
(W_{opt})_{MFLD} = \arg \max_w \frac{|W^T S_w W|}{|W^T S W|}
\]  

(1.3)

This method was proposed to solve SSS problem. It has been shown that the original Fisher’s Linear Discriminant Criterion can be replaced by the modified Fisher’s Linear Discriminant (MFLD) criterion in the course of solving the discriminant vectors of the optimal set [29]. In this method, all image samples are first projected onto the null space of \( S_w \) and then, PCA technique is applied on the projected samples to obtain the optimal projection vectors. Chen et al. also proved that by applying this method, the modified Fisher’s Linear Discriminant criterion attains its maximum. Therefore, the null space method extracts features which are optimal from a discrimination point of view. It turns out that resulting orthonormal projection vectors span the space obtained as the intersection of the null space of \( S_w \) and the range space of \( S_r \). This space is called as optimal discriminant subspace
since it is spanned by vectors that extract optimal features for discrimination. However, the algorithm proposed by Chen et al. is not an efficient algorithm for application in the original sample space. This method reduces the dimension of the original sample space and then the null space method is applied in the reduced space. Moreover H. Cevikalp et al. [31] observed that the performance of the null space method depends on the dimension of the null space of $S_w$ in the sense that larger dimension provides better performance.

Huang et al. [27] proposed the PCA+NULL Space method to solve the SSS problem. In this method, PCA is applied to remove the null space of $S_r$, which contains the intersection of the null spaces of $S_b$ and $S_w$. Then, the optimal projection vectors are found in the remaining lower dimensional space by using the null space method. In this method, the within class scatter matrix in the reduced space is typically singular. Yang et al. [30] applied a variation in this method to make $S_w$ as non-singular. For this, after dimension reduction, they split the new within class scatter matrix into its null space and orthogonal complement. Then, all projection vectors that maximize the between class scatter in the null space are chosen. However, applying PCA and using all eigen vectors corresponding to the nonzero eigen values make these methods impractical for face recognition applications when the training set size is large. This is due to the computational expense of training being very large. Yu. H et al. [25] proposed the Direct LDA method which uses the simultaneous diagonalization method in [22]. In this method, the null space of $S_b$ is removed and, then the projection vectors that minimize the within class scatter in the transformed space are selected from the range space of $S_b$. However, removing the null space of $S_b$ by dimensionality reduction will also remove part of the null space of $S_w$ and may result in the loss of important discriminative information [26, 27].

Cevikalp et al. [31] proposed the Discriminative Common Vector (DCV) method for finding the optimal orthonormal projection vectors in the optimal discriminant
subspace. This method is equivalent to the null space method, with the exception that the dimension reduction step is omitted and therefore this method exploits the original high dimensional space. Two efficient algorithms were proposed to compute the optimal projection vectors. One of the algorithms uses the range space of $S_w$, while the other uses subspace methods and the Gram-Schmidt orthogonalization procedure. This method can be used only when the dimension of the sample space is larger than the rank of $S_w$. i.e., in SSS case. This method achieved the better performance than other methods in terms of accuracy, real-time performance, and numerical stability. However, the DCV method is still a linear technique in nature, which is inadequate to describe the complexity of face image due to illumination, facial expression and pose variations.

Discriminant analysis techniques utilizing kernels have been proposed in the work reported in [32-34] to extract non linear features of face image. In this work the input data are transformed into a higher dimensional space by a nonlinear mapping function and then apply the linear discriminant analysis techniques in that space. These methods are formulated in terms of dot products of the mapped samples, and kernel functions are used to compute these dot products. Therefore, the non-linear mapping function and the mapped samples are not used explicitly, which makes the methods computationally feasible. However, the singularity problem of the involved matrices is typically encountered in this approach since the dimensionality of the mapped space is usually larger than the size of the training set.

Many techniques have been adopted to solve the SSS problem in kernel approaches. Mika et al. used the original FLD criterion in the nonlinearly mapped space and added a small perturbation matrix to the involved singular matrix [32]. This approach is called as Kernel Fisher’s Discriminant Analysis (Kernel FDA). Yang et al. [34] proposed the kernel PCA + LDA method. In this method, Modified FLD criterion is used instead of the original FLD criterion. The procedure followed in this method is that the training set samples are projected onto the range space of
total scatter matrix using the kernel PCA method [35], then applied the LDA method in the reduced space which maximizes the modified FLD criterion.

Ceviaiklp et al. [36] proposed the Kernel Discriminative Common Vector (KDCV) method. In this method, the input samples are transformed into a higher dimensional space by a nonlinear mapping function and then applied the DCV method in the nonlinearily transformed higher dimensional space to find the optimal orthonormal projection vectors in the optimal discriminant subspace. They used the modified FLD criterion in the mapped space instead of the original FLDA criterion to attain its maximum value. This method improved the recognition results on different face databases having nonlinear and complex distributions. They proved that the real time performance of KDCV method is superior to other kernel methods like Kernel PCA and Kernel LDA and also in solving SSS problem encountered in the transformed space.

The above mentioned methods are suitable for any face recognition applications in SSS case. However, in the case of Large Sample Size (LSS) i.e. the number of samples in the training set being larger than the dimension of the sample space, the recognition performance of these methods is reduced. These methods are not adequate for an application like Voter's identification using their face images since these methods are developed especially for SSS case. Therefore the later research on pattern recognition aimed to develop face recognition methods for LSS case.

Statistical and Probabilistic approaches are commonly used in classification problems. One of the statistical methods is based on subspace methods. The most well known subspace method is Fisher's Linear Discriminant Analysis (FLDA) method. FLDA is an important method for linear dimensionality reduction in statistical pattern classification with SSS and LSS, and applied this method for many applications like word recognition; face recognition etc., [37-39]. Although FLDA is the linear transformation that maximizes the mean squared distance between the classes in lower dimensional feature space, it is not optimal in respect
to minimizing classification error rate in that space. Loog et al. proposed a
generalized version of FLDA that allows de-emphasizing of the contributions of
classes far apart from each other [40]. In this criterion, the differences among class
means are also considered by means of defining a generalized between class scatter
matrix.

Yang et al. emphasized that the Fisher criterion is not an absolute criterion and it
should be associated with statistical correlation to assess the discrimination of a set
of discriminant vectors [41]. Gulmezoglu et al. proposed the Common Vector (CV)
subspace classifier method that eliminates unwanted information, such as
environmental effects, personal and phase differences, and temporal variations from
a spoken word [42]. In this method, the subspaces are modeled with their basis
vectors spanning the null space of the covariance matrix of each class. This method
is based on the assumption that the number of samples in each class is smaller than
or equal to the dimensionality of the sample space [42, 43] i.e. SSS. It has been
successfully applied in the isolated word recognition, face recognition, speaker
recognition applications [42-45].

Gulmezoglu et al. developed the Common Vector (CV) method for LSS case also
[46]. In LSS case, the number of samples in the training set is sufficient to calculate
the inverse of the within class scatter matrix. This inverse does not exist for SSS
case. The common vector will be the zero vectors in the LSS case because
indifference subspace disappears or none of the eigen values of the within class
scatter matrix is going to be zero in the LSS case. Therefore, in this method,
Gulmezoglu, showed that the indifference subspace will not disappear as long as
some of the eigen values are very small when compared with others in the within-
class scatter matrix. Then they obtained the common vector of each class by
projecting the class mean vector onto the indifference subspace. Moreover, they
showed that this method extracts temporal variations as well as all the other
differences and better recognition rate is achieved than the other subspace methods.
It was demonstrated that if the training set samples are linearly independent, then
the extracted features are optimal and all training set samples can be classified correctly [42,43].

Gulmezoglu et al. [43] and Cevikalp et al. [36] proved that the projections of all samples onto the null space of the total scatter matrix gives rise to the same vector i.e., this subspace does not contain any discriminative information for classification of samples. Then they proposed the new subspace for representing each class and can be defined as the intersection of the null space of that class’ scatter matrix and the range space of the total scatter matrix. Cevikalp et al. proposed Common Vector method based on this new subspace to represent classes and named as Modified Common Vector (MCV) method in [47]. In this method, a class subspace is represented by using the basis vectors spanning the intersection of the null space of that class’ scatter matrix and the range space of the total scatter matrix. They showed that the recognition accuracy of MCV method is the same as CV method and this method is inadequate to extract non linear features of face images. However, this method may not extract nonlinear features of classes since each class is associated with a linear subspace [48, 49].

Suda [49] proposed the kernel-based nonlinear subspace method which is called the Kernel Class-featuring Information Compression (Kernel CLAFIC) method to overcome this limitation. In this method, it is assumed that samples from each class lie in some nonlinear subspace. Therefore, all samples in each class are mapped into a higher dimensional feature space through a nonlinear kernel mapping function, and the Kernel PCA [50] has been employed to compute the principle components of the correlation matrices of classes in the mapped space. This Kernel CLAFIC method is formulated in terms of dot products of the mapped samples, and kernel functions are used to compute these dot products. Therefore, the mapping function and the mapped samples are not used explicitly, which makes the method feasible.

Cevikalp et al. proposed Kernel Common Vector (KCV) method to overcome the limitations of MCV method [47]. In this method, the training set samples are
mapped into an implicit higher dimensional space using a non linear mapping function and then applied the MCV method in the mapped space to extract the common vector of each class. By using the kernel trick, the CV method is formulated in terms of dot products of the mapped samples, which are computed using kernel functions. As a result, the mapping function and the mapped samples are not used explicitly, which makes the method computationally feasible. They proved that the recognition rate of KCV method is superior to the Kernel CLAFIC method.

From the above literature survey, it is observed that the KDCV method improved the recognition accuracy in SSS case and the KCV method improved the recognition accuracy in LSS case. These methods are promising methods for feature extraction in the sense that these methods employ a feature space such that the class's separability is maximized. Moreover these methods are suitable for recognition of face images varying in facial expressions and illuminative conditions. The recognition performance of these methods depends on the clustering of samples into classes in the training set. If there is any mislabel in the training set, these methods fail to classify a given test sample and these methods act like a weak classifier in this environment. This will be solved using boosting technique. Thus the survey of boosting techniques is required and is as follows.

Schapire [51] showed that any weak learning algorithm can be efficiently transformed or boosted into a strong learning algorithm. Later, Freund [52], [53] has presented, “boost-by-majority” algorithm that is considerably more efficient than Schapire’s algorithm. Both algorithms have been worked by calling a given weak learning algorithm multiple times, each time presenting it with a different distribution over the input domain and finally combining the entire generated hypothesis into a single hypothesis. The main idea is to alter the distribution over the input domain in a way that increased the probability of the harder samples in the sample space, thus forcing the weak learner to generate new hypothesis thus make
less mistakes on these harder samples. However, these algorithms cannot take advantage of hypothesis computed by weak learning algorithm.

Freund and Schapire [54] have proposed a new boosting algorithm, AdaBoost algorithm which is very nearly as efficient as boost-by-majority. However, unlike boost-by-majority, the accuracy of the final hypothesis produced by this new algorithm depends on the accuracy of all the hypotheses returned by weak learning algorithm. That is, it adjusts adaptively to the errors of the weak hypothesis returned by weak learning algorithm Drucker et al. [55] have performed experiments using boosting to improve the performance of a real valued neural network, observed that summing the outcomes of the networks and then selecting the best prediction performs better than selecting the prediction of each network and then combining them with a majority rule. The AdaBoost algorithm has used the combination rule that was observed to be better in practice. Several successful experiments have been conducted using AdaBoost, including work by Freund and Schapire [56], Drucker et al. [57], Kita [58], and Boris Ruf [59].

In the AdaBoost algorithm, the dependency of classifier output is maintained to be low by providing training samples with different weights that are increased based on the difficulty in classification [54]. To ensure low dependency of classifier outputs, Lu et al. [60] have proposed an interesting approach in which at each boosting round a new LDA is performed to construct a low-dimensional feature space called LDA subspace by focusing on hard-to-separate training samples, and the features projected to this LDA subspace are trained by a weak learner model. S.Kita et al. extended the above Lu’s work [60] in which KDA is adopted instead of LDA as a feature extraction method [61]. They have called this approach as “Boosting Kernel Discriminant Analysis (BKDA). In BKDA, to satisfy the weak learner condition, a small number of training samples are selected based on a weight function and they are applied to KDA. Training such a small training set also reduces the computation costs of KDA even when many classifiers are created. The feature extraction by Kernel Discriminant Analysis (KDA) is first carried out to
construct a feature space for every classifier, and then the feature vectors are used for training the classifier. Kita et al. [61] have improved their prior work in [58] by means of selecting the kernel function within a feasible time and by the criterion of the training convergence. Adaboost is a type of boosting algorithm which works by repeatedly applying a given weak learner to a weighted version of the training set and then linearly combined these weak classifiers constructed in each iteration into a single strong classifier. It reduces the generalization error in many real-world applications. For multi-class problems, Adaboost.M2 outperforms Adaboost.M1. The method proposed in [58] is a multiple classifier system by combining Adaboost.M2 and KDA. When applying KDA, the training samples are selected based on the distribution probability which is also used for selecting samples to train a classifier in Adaboost.M2. This sample selection leads not only to the reduction of the computation costs but also to make every classifier uncorrelated, resulting in boosting the performance of Adaboost.M2. They showed that the performance of BKDA method is superior to other methods like Boosting LDA [62] and Support Vector Machine (SVM) with regard to the recognition rate. Thus the AdaBoost.M2 boosting technique improves the recognition performance of classification methods by constructing strong classifier.

1.3 CURRENT STATE OF THE ART

1.3.1 Kernel Discriminative Common Vector (KDCV) Method

In the last decade, many methods have been proposed to solve the SSS problem in face recognition process. Among these, Discriminative Common Vector (DCV) method proposed by Cevikalp et al. [31] successfully overcomes the drawback of other methods for solving SSS problem. This method is equivalent to the null space method, with the exception that the dimension reduction step is omitted and therefore this method exploits the original high-dimensional space. Two efficient algorithms were given to compute the optimal projection vectors in the optimal discriminant subspace. One algorithm uses the range space of within-class scatter matrix while the other uses subspace methods and the Gram-Schmidt
orthogonalization procedure. The optimal projection matrix $W$ is obtained by using any one of the above mentioned algorithms which avoids solving the eigen equation as in other methods. Therefore in the DCV method the computational complexity is significantly reduced and the stability of computation is improved.

However, the DCV method is still a linear technique, which is inadequate to describe the complexity of face image due to illumination, facial expression and pose variations. Therefore, discriminant analysis techniques utilizing kernels have been proposed in [32-34]. Their main idea is to transform the training samples into a higher dimensional space by a nonlinear mapping function and then applied the linear discriminant analysis techniques in that space. These methods are formulated in terms of dot products of the mapped samples, and kernel functions are used to compute these dot products. Therefore, the nonlinear mapping function and the mapped samples are not used explicitly, which makes the methods computationally feasible. Since the mapped space is arbitrarily large dimensionality, the SSS problem is encountered in these approaches. Cevikalp et al. [36] proposed Kernel Discriminative Common Vector (KDCV) method to solve the SSS problem encountered in kernel approaches and to overcome the limitation of DCV method.

In KDCV method, the training set samples are mapped in to a higher dimensional space by a nonlinear mapping function and then DCV method is applied to extract optimal discriminant features. The modified FLDA criterion is not appropriate since the maximization does not have a unique solution in the SSS case. In particular, every projection vector matrix $W$ such that $W^T S_\omega W = 0$ and $W^T S_y W \neq 0$ maximizes the modified FLDA criterion. Note that if $S_\omega$ is singular, which is always the case for the SSS problem, there are many such matrices $W$. However it is not reasonable to use matrices $W$ with a small number of projection vectors since they may not be sufficient for optimal feature extraction. On the other hand, the following criterion, called the null space based FLDA criterion (NFLD), has a unique maximum for the projection vectors with unit length and also maximizes the modified FLDA criterion:
\[(W_{opt})_{NFDL} = \arg \max_{W^T S_y W = 0} |W^T S_y W| = \arg \max_{W^T S_y W = 0} |W^T S_y W|. \] (1.4)

The optimal projection vectors maximizing this criterion are obtained by projecting the training set samples onto the null space of within-class scatter \(N(S_w)\) and then obtained the projection vectors by performing PCA technique. As a result, a set of orthonormal vectors that forms a basis for the space which is called optimal discriminant subspace is obtained. This subspace is defined as the intersection of \(N(S_w)\) and the range space of total scatter matrix \(R(S_i)\). The criterion given in Equation (1.4) attains its maximum for orthonormal vectors that form a basis for the optimal discriminant subspace.

In this method, the training set samples are mapped into a higher dimensional feature space as \(\phi: R^d \rightarrow F\) and the within-class scatter matrix \(S_w^\phi\), the between-class scatter matrix \(S_b^\phi\) and the total scatter matrix \(S_t^\phi\) are defined as

\[S_w^\phi = \sum_{i=1}^C \sum_{k=1}^{N_i} (\phi(x_i^k) - \mu_i^\phi)(\phi(x_i^k) - \mu_i^\phi)^T\] (1.5)

\[S_b^\phi = \sum_{i=1}^C N_i (\mu_i^\phi - \mu^\phi)(\mu_i^\phi - \mu^\phi)^T\] (1.6)

and

\[S_t^\phi = \sum_{i=1}^C \sum_{k=1}^{N_i} (\phi(x_i^k) - \mu^\phi)(\phi(x_i^k) - \mu^\phi)^T = S_b^\phi + S_w^\phi\] (1.7)

where \(\mu^\phi\) is the mean of all samples and \(\mu_i^\phi\) is the mean of samples in the \(i^{th}\) class.

In the transformed space \(F\), \(S_w^\phi\) is typically singular due to the high dimensionality of the mapped space. Thus the optimal projection vectors that maximize the null space based LDA criterion are in the intersection of the null space \(N(S_w^\phi)\) of \(S_w^\phi\) and the range space \(R(S_t^\phi)\) of \(S_t^\phi\). First, they projected the training set samples onto \(R(S_t^\phi)\) through the kernel PCA. Then they found the vectors that span the new null space of the within-class scatter matrix of the transformed samples. Consequently, the discriminative common vector that represents each class is obtained.
This method has extracted features of face images varying due to illumination and facial expressions by mapping samples to a higher dimensional space via nonlinear mapping function. Moreover, this method has correctly classified the training samples by using the optimal projection matrix W. Thus the performance of recognition is increased. However, the goal of a recognition method is not only to classify all training samples but also to classify well the test samples that are not used for training. In other words, the recognition method has to produce a correct input-output mapping. This is known as the generalization ability of a method [63]. Cevikalp et al. proved that this method possesses high generalization ability. The dimensionality of the sample space and the size of the training set are two important factors that affect the performance of recognition methods [64]. Cevikalp et al. conducted experiments on data sets from two different populations with different training set sizes and dimensionalities [36]. They have tested their KDCV method on Fisher's Iris database and the Digit dataset of handwritten numerals which have large number of samples in the training set. Their test results show that this method generalizes well compared to other kernel approaches for data sets with large number of samples. They have tested the KDCV method with Olivetti-Oracle Research Lab (ORL) face database and showed that the KDCV method outperforms other kernel methods and the recognition rate is increased over linear DCV method. They concluded that the KDCV method provides a reliable input-output mapping for the data sets with high dimensional space by using only a few training set samples. Additionally, the KDCV method improved the recognition result of the linear DCV method on different face databases having nonlinear and complex distributions. Another advantage of KDCV method is its real time performance. Because, in this method, after a test image is projected onto the (C-1) optimal projection vectors, the feature vector of the test sample is compared to C discriminative common vectors only. Whereas in other methods, the feature vector of a test sample is compared to all the training set feature vectors.
1.3.2 Kernel Common Vector Method (KCV)

Gulmezoglu et al. proposed the Common Vector (CV) subspace classifier method that extracts features that are common to all samples in each class [43]. The common features of each class are extracted by eliminating the features that are in the direction of the eigenvectors corresponding to the nonzero eigen values of scatter matrices of classes. In this way, each class is represented by the null space of its own scatter matrix. Therefore this method can be applied in SSS case only. Gulmezolu et al. [46] developed the CV method for LSS case. In this method, the common vector of each class is obtained by projecting the class mean vector onto the indifference subspace. The drawback of CV method is, it cannot be extended to the nonlinear case directly since this method cannot be formulated using dot product of the mapped samples. Moreover Cevikalp et al. [36] showed that in CV method the null space of the total scatter matrix does not contain any discriminative information for classification of training samples. Therefore this subspace is discarded to represent a class subspace. Cevikalp et al. proposed the Modified Common Vector (MCV) method to overcome these drawbacks [47] and defined a new subspace for representing each class as the intersection of the null space of that class’ scatter matrix and the range space of the total scatter matrix. In this method, a class subspace is represented by the basis vectors spanning these intersections. They showed that the recognition accuracy of MCV method is same as CV method. However, the MCV method can be extended to nonlinear case. Cevikalp [47] introduced a new nonlinear subspace classifier by incorporating the kernel trick into the MCV method and this method is called as Kernel Common Vector (KCV) method.

In KCV method, the training set samples are mapped into an implicit higher dimensional space \( F \) using a nonlinear mapping function and then applying the linear MCV method in the mapped space. As in other kernel methods using kernel trick, the KCV method is also formulated in terms of the dot products of the mapped samples, which are computed using kernel functions. As a result, the mapping function and the mapped samples are not used explicitly, which makes the
method computationally feasible. In this method, the scatter matrix $S_{i}^{\phi}$ of each class and the total scatter matrix $S_{i}^{\Phi}$ are defined as

$$S_{i}^{\phi} = \sum_{k=1}^{N_{i}} (\phi(x_{i}^{k}) - \mu_{i}^{\phi})(\phi(x_{i}^{k}) - \mu_{i}^{\phi})^{T} \quad (1.8)$$

$$S_{i}^{\Phi} = \sum_{i=1}^{C} \sum_{k=1}^{N_{i}} (\phi(x_{i}^{k}) - \mu^{\phi})(\phi(x_{i}^{k}) - \mu^{\phi})^{T} \quad (1.9)$$

where $\mu_{i}^{\phi}$ is the mean of mapped samples in the $i^{th}$ class, and $\mu^{\phi}$ is the mean of all mapped samples. In this method, a class subspace is defined by the basis vectors of the intersection subspaces $N(S_{i}^{\phi}) \cap R(S_{i}^{\phi}), i = 1, \ldots, C$. To find these basis vectors, they transformed all training samples onto $R(S_{i}^{\phi})$ and then found the vectors spanning the null spaces of scatter matrix of each class. Consequently the common vector that represents each class is obtained.

The ratio of the dimensionality of the sample space to the training set size is a very important factor that affects recognition performances of subspace classifiers. Therefore, this method is applied on seven different real world data sets having various ratios of training set sizes to dimensionalities. They have applied this method on AR face database and achieved the higher recognition rates among all other subspace classifier methods. They showed that when the number of samples in a class is increased, the generalization performance of this method is also improved. From the obtained results on different data sets, they have observed that the generalization performance of KCV method depends on the dimensionality of the sample space and gives better recognition rates in that higher dimensional sample space i.e., in the case of SSS. Another given observation from their experimental results is that the intersection of the range space of the total scatter matrix and the null space of each individual class scatter matrix is typically the best subspace for discrimination. This method can also be used when the number of samples in each class is larger than the dimensionality of the sample space using kernel functions i.e., in the case of LSS.
1.4 MOTIVATION FOR THIS RESEARCH

Limitations of KDCV Method

The previous section describes about the KDCV method which is a feature extraction cum classifier method. This KDCV method improves the recognition results of the linear DCV method on different face databases having nonlinear and complex distributions. The real time performance of KDCV is superior to other kernel method like Kernel PCA and Kernel LDA. However the recognition accuracy of KDCV method is reduced due to the following limitations.

- The expression of the between-class scatter matrix $S_w^o$ measures the distribution of mean of each class from the centre mean. In this expression, the overlapping between the classes is not considered.
- The within-class scatter matrix $S_w^o$ assumes that all classes have the same weight for its covariance. If there is any class with dominant covariance then the within-class scatter matrix $S_w^o$ will fail to estimate the correct value for classification.
- The algorithm that followed to obtain discriminative common vectors is computationally expensive and numerically unstable.

The above mentioned limitations of KDCV method can be solved as follows:

- Modify the between-class scatter matrix such that this matrix considers the pair wise discriminatory information between the classes.
- Modify the within-class scatter matrix such that it assigns weight for class’ covariance based on its contents.
- Obtain discriminative common vectors by using the Difference subspace and Gram-Schmidt orthogonalization procedure which is numerically stable and reduces computational complexity.

Limitations of KCV Method

Cevikalp et al. [47] proved that the recognition performance of KCV method is superior to other kernel subspace methods in the SSS case. In addition, they showed that this method can also be applied, when the number of samples in the training set
is larger than the dimensionality of sample space i.e. in LSS case. However, the performance of recognition in KCV method is reduced due to the following limitations.

- In this method, classes are modeled as separate subspaces in the sample space. In some cases, a class’ subspace may interfere with other class’ subspace i.e., there is overlapping between the classes in the sample space. This will reduce the performance of the recognition process.

- KCV method considered only the class’ scatter matrix and the total scatter matrix. Since there is no consideration of between-class scatter matrix in this method, the overlapping between class’ subspaces will be ignored. Thus, it becomes necessary to consider the between class scatter matrix in the method.

These limitations can be solved by introducing a between class scatter matrix. Since, classes are modeled as separate subspaces, this new between-class scatter matrix has to consider the pair wise discriminatory information between the classes.

From the literature survey, it is observed that the KDCV and KCV methods improved the recognition accuracy in SSS case and the KCV method can be applied in LSS case. However, the recognition performance of these methods depends on the clustering of samples into classes in the training set. Moreover, if there is any mislabel in the training set, these methods fail to classify a given test sample and these methods act like a weak classifier in this environment. To overcome this, some other technique is to be included in the method such that all the samples in the training set are labeled correctly which in turn convert KDCV and KCV based weak classifiers as strong classifiers. This may increase the recognition accuracy on test samples.

These limitations of KDCV and KCV methods have motivated to develop improved KDCV and improved KCV methods overcoming some of the above said limitations.