Chapter 2

Principal Component Analysis Methods: A Literature Survey

2.1 Introduction

Note: The work in this chapter has been submitted to *Journal of Pattern Recognition Research*[1].

In this Chapter we review the literature related to Principal Component Analysis (PCA) methods in brief. For better understanding we classify the literature (Figs. 2.1 to 2.4) into various categories viz, Feature Partitioning or Block based PCA (FP-PCA) methods, 2D structure based PCA methods, Artificial Neural Network based methods, Kernel PCA methods, EM algorithms to PCA, Hybrid methods, Choosing number of Principal Components, etc., as described in the following sections. In

[1]Kadappagari Vijaya Kumar and Atul Negi, "A review of principal component analysis methods", **submitted to** *Journal of Pattern Recognition Research*, on Jan. 13^{th} 2009.

addition, we discuss how the existing PCA methods solve the problems of classical PCA. At the end, we state the problem we address in our investigations in this thesis.

2.2 Feature Partitioning or Block based PCA (FP-PCA) Methods

It is now known that classical PCA suffers from the drawbacks of not coping well with high dimensional data and scaling up to large data set due to its prohibitive computational complexity ($O(N.d^2)$). Another shortcoming is that classical PCA may not perform well in terms of recognition for applications where local region based features have discriminant information (e.g. facial expressions, pose, illuminations, etc and change detection applications). To overcome these problems, Block-based PCA methods (henceforth we call them as feature partitioning based PCA (FP-PCA) methods) were emerged. Chen et al proposed Sub-pattern based PCA (SubPCA) technique [21] which divides each pattern into equally-sized sub-patterns and groups similar sub-patterns from all patterns into corresponding sub-pattern set. Local features are extracted from each sub-pattern set and are concatenated to form reduced patterns. Chen et al proved that SubPCA [21] outperforms PCA in terms of classification. Other approaches that appear similar to SubPCA method are Multi-Block PCA [128] and Region-based PCA [136]. Multi-Block PCA [128] is used for change detection in remote sensing. Region-based PCA [136] is used in clutter rejection technique for Forward Looking Infra Red (FLIR) imagery in Automated Target Detection application. Region-based PCA technique is proposed to categorize all target images by

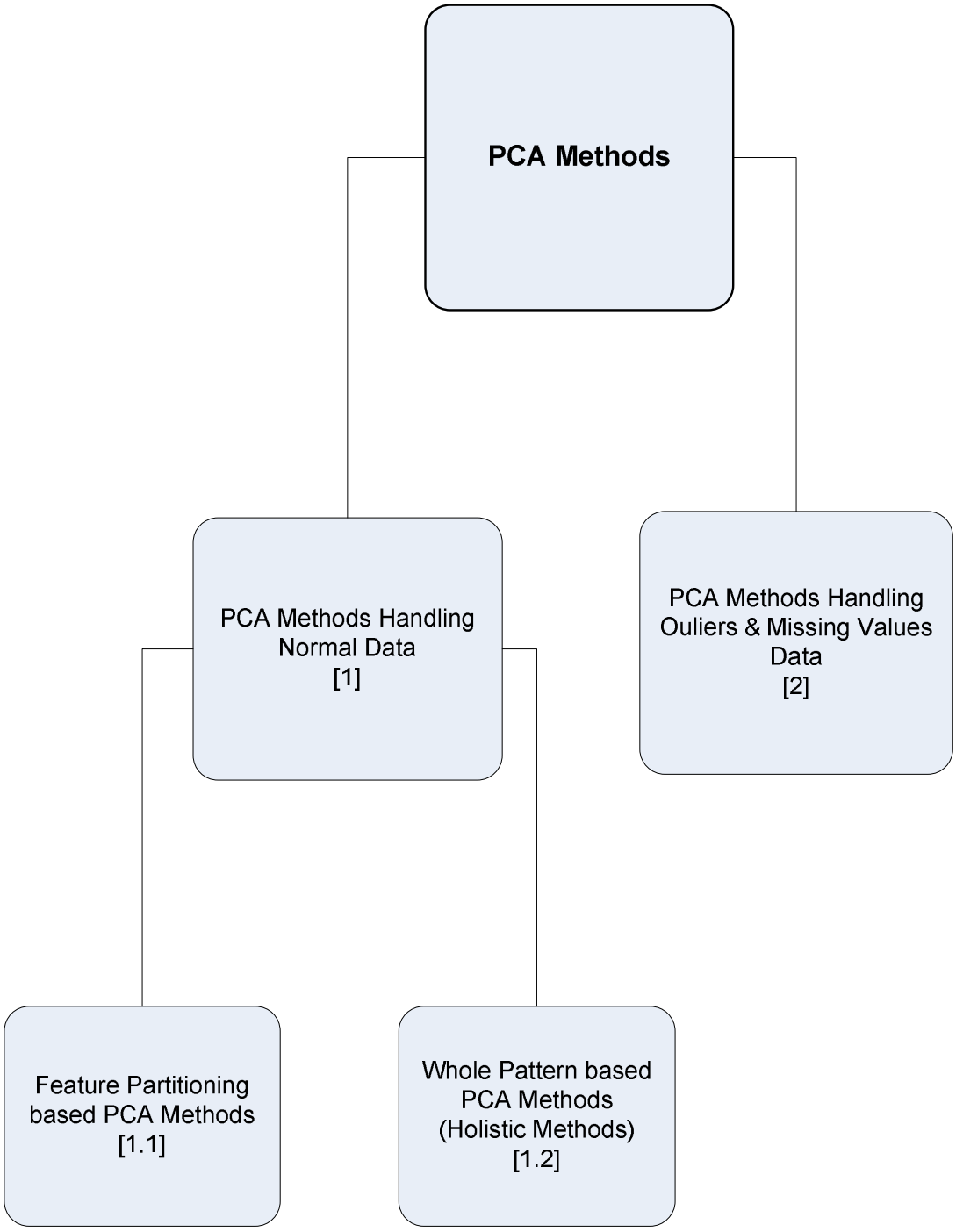

Figure 2.1: Main classification diagram of PCA methods

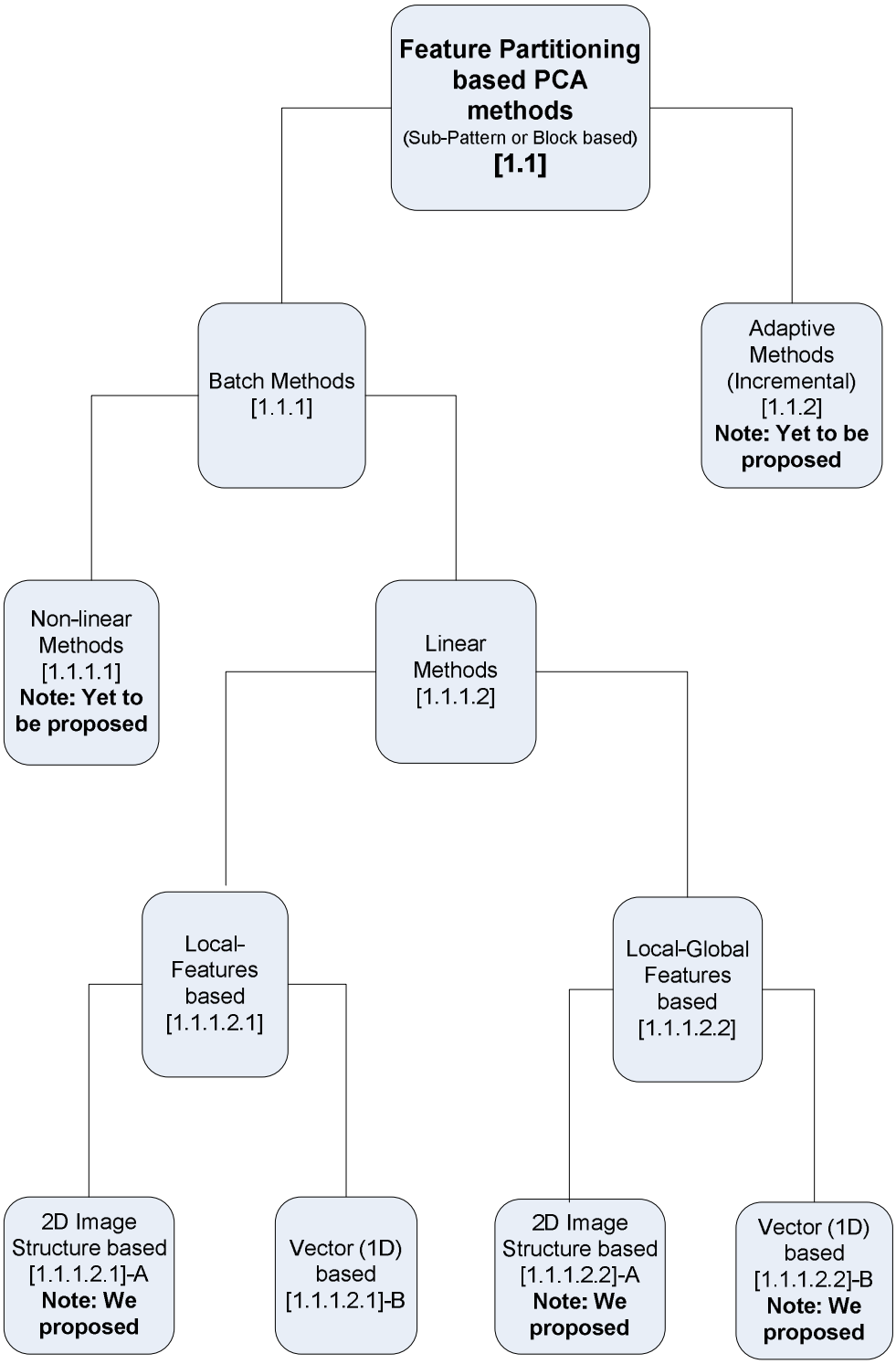

Figure 2.2: Classification of FP-PCA (sub-pattern based PCA) methods

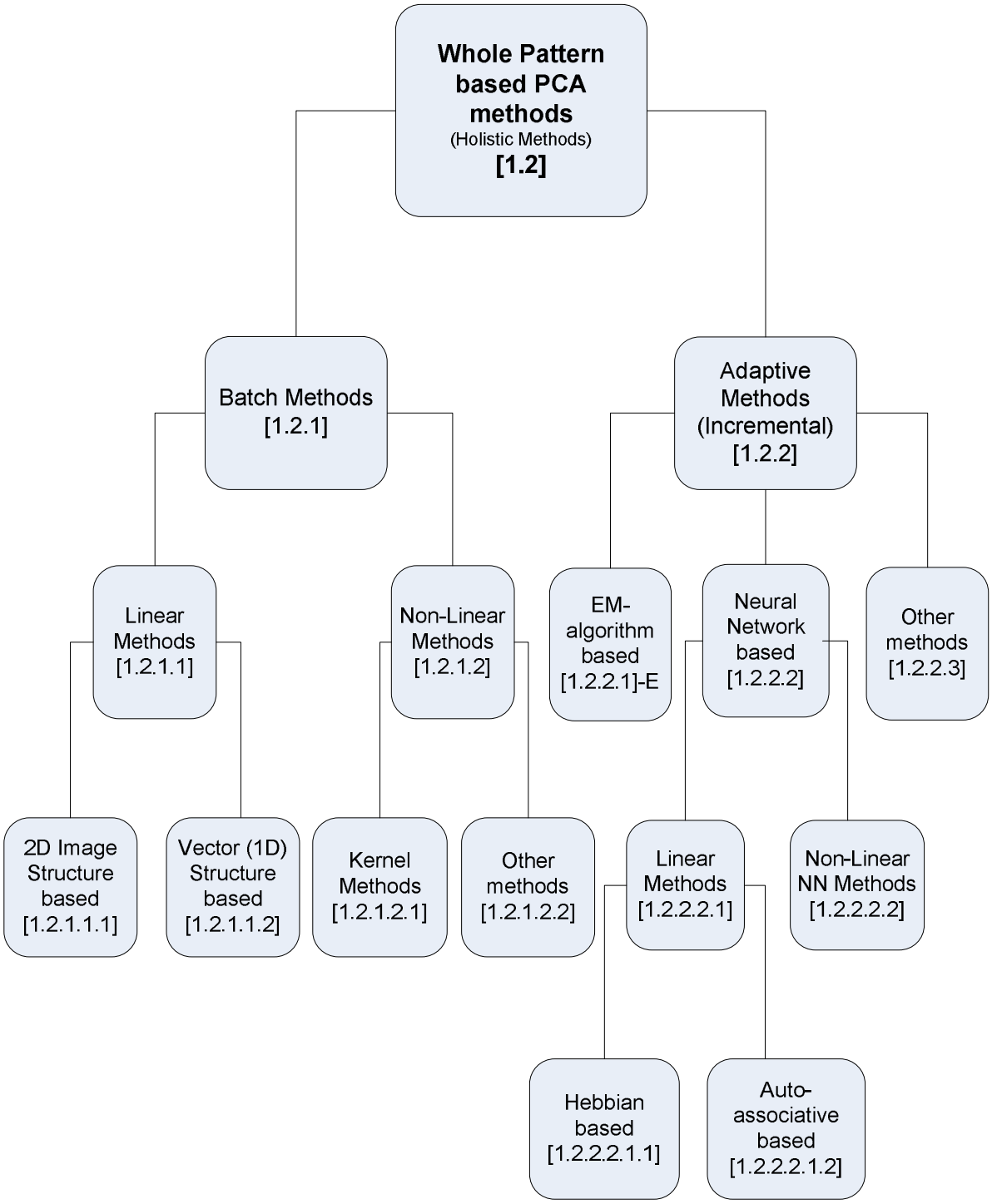

Figure 2.3: Classification of whole-pattern based PCA (global PCA) methods

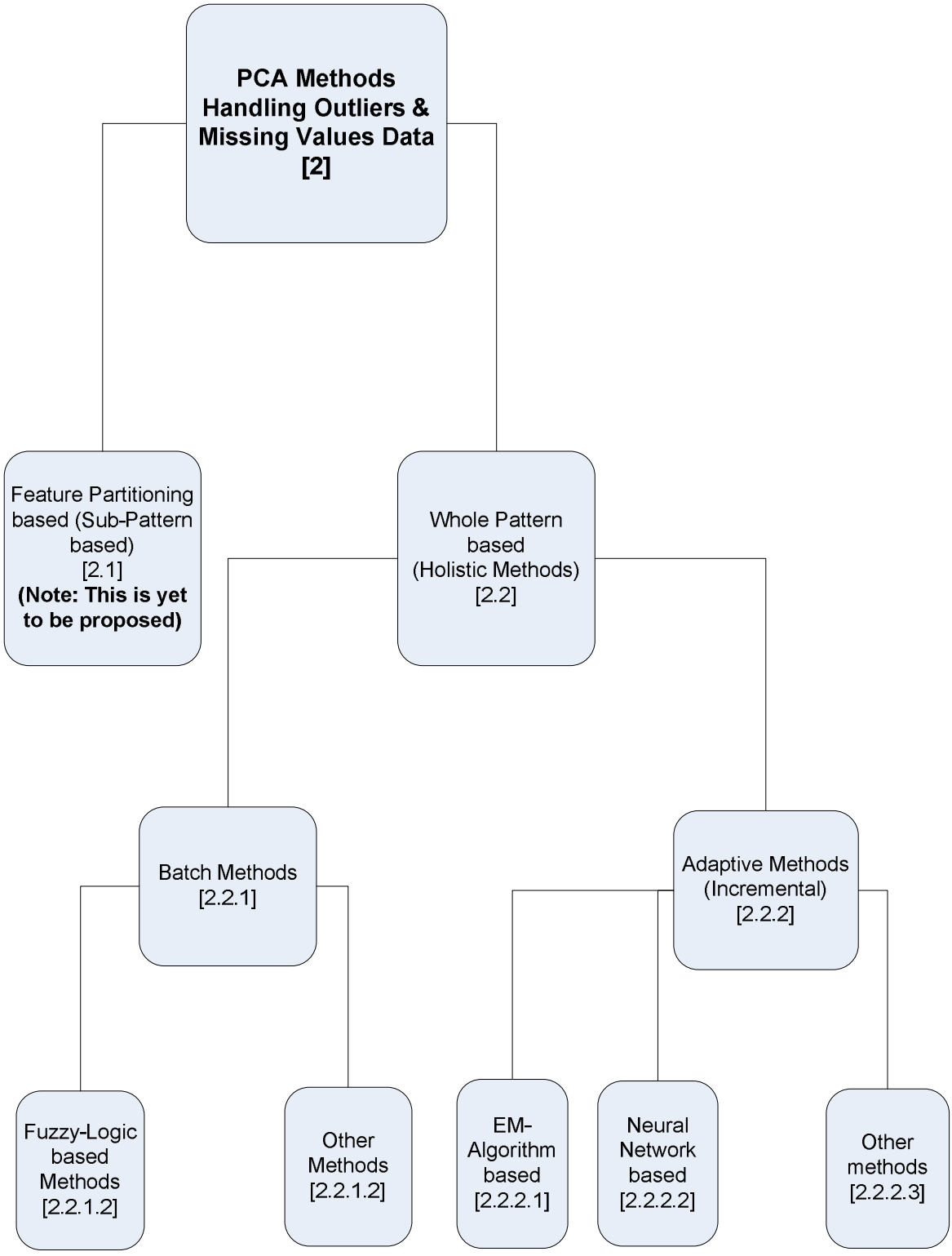

Figure 2.4: Classification of PCA methods for outliers and missing values data

clustering together the target images with respect to their similar sizes and shapes in order to form a group. Each group is further divided into several regions, and a PCA is performed for each region in a particular group to extract feature vectors.

Some improvements in this direction which combine the local features more structurally include Localized PCA [106] [107], Aw-SpPCA [157] and Clustered Block-wise PCA [113]. In localized PCA [106] [107], first they localize faces using skin colour. Erosion and dilation operations from mathematical morphology are used to get rid of small isolated segments. Then a PCA technique is used to localize mouth and eyes. Face is rotated until frontal position is obtained using x and y coordinates of two eyes, followed by contour technique to search for boundaries of the face (i.e. top, left, down, right). Then it is morphed to get a standard face. Next the faces are divided into k regions. Then PCA is applied for each region similar to SubPCA-like methods to extract features. Local features are combined using Probabilistic approach (adding local probability densities). Localized PCA focus more on localizing face image using different image processing techniques and needs a lot of preprocessing of images. Another method, Adaptively Weighted SubPCA (Aw-SpPCA) [157] operates directly on its sub-patterns partitioned from an original whole pattern and separately extracts local features from them. Moreover, Aw-SpPCA can adaptively compute the contributions of each part and then uses them to a classification task. However, Aw-SpPCA has additional burden of computing contributions. Clustered Block-wise PCA [113] uses an algorithm developed by Hall et al [56] to merge a pair of subspaces. For merging subspaces raw data is not required. For each block, distance is computed between its subspace and all other subspaces and store the block number

whose subspace fall below a distance threshold. After listing all the subspaces that are close to each other in terms of subspace distance, each pair is merged one-by-one with the subspace merging algorithm. Clustering and merging the block subspaces results in a reduction in the number of subspaces. When a new set of images are added to the data, once the number of these images becomes equal to the temporal size of the block, one can apply PCA to the blocks within this new data set and merge the subspaces that are close to existing subspaces. In this way, the necessary storage will not increase linearly to the size of the added data and correlation of local visual events can be exploited as new data is added. If an existing subspace is merged with a subspace computed from a new data block, the projection of the existing data block should be updated with the projection in the newly merged subspace. However Clustered Block-wise PCA suffers from the drawbacks: (i) it has quadratic (in terms of number of blocks) time complexity and may be prohibitive if the number of blocks is high and (ii) it has overhead to update projection for each merge.

A slightly different approach, Sub-Holistic PCA (SHPCA) [80] was proposed for face recognition. In this scheme, each face image is not only taken as a whole but four equally-sized sub-images are also formed from the given image. In this scheme, instead of generating a single face space, five face spaces are generated. The image to be tested is also divided into four parts and the complete image with the four sub-parts is projected in their respective face spaces. The results from all five face spaces are obtained and from the five proposed matches one match is found. However, SHPCA suffers from the following drawbacks: (i) It needs to compute original face space, in addition to 4 new face subspaces, which is computationally intensive and (ii)

it divides each face into 4 sub-patterns only, which may not be correct for all faces.

The methods discussed so far in this section have a similarity: *they divide each pattern into sub-patterns and apply classical PCA to each of the sub-patterns separately.* We call these methods as SubPCA-like methods because they perform feature extraction in the similar way as SubPCA method.

In contrast to SubPCA and similar approaches, modular PCA approach (modPCA) [53] divides each pattern into sub-patterns and *apply single PCA to the set of all sub-patterns to find principal eigenvectors, instead of applying single PCA to each of sub-patterns.* Then, the sub-patterns are projected onto the same principal eigenvectors to extract local features. Another approach similar to modPCA, called Eigen-regions method [45], uses segmentation techniques to divide the given image into meaningful regions, and then applies single PCA to all these regions. Eigen-regions method uses down-sampling procedure to reduce the size of a region, so that PCA can be applied to reduce computation.

In contrast to other PCA methods which are based on whole patterns, FP-PCA methods are unique in their approach, which are based on novel idea of *partitioning each pattern into sub-patterns and extract local features.* FP-PCA methods alleviate some of the crucial problems of classical PCA and show (i) Reduced computational complexity, (ii) improved recognition/classification rate by local feature extraction if local variations are prominent, etc. However, FP-PCA methods suffer from the following problems: (i) These methods purely perform local feature extraction, that is feature extraction is limited to a subset of original feature set (or limited to sub-patterns). Therefore, FP-PCA methods may not perform well if there exists global

variations, (ii) Summarization of variance is not good because the entire covariance structure is not utilized which yields more number of local PCs resulting in less dimensionality reduction, (iii) FP-PCA methods do not exploit two-dimensional structure of image data (Section 2.3) because they all use classical PCA in a region or block to extract local features.

2.3 Two Dimensional Image Structure based Methods (2DPCA and Its Variants)

PCA and its disadvantages to image data:

One of the most successful image recognition applications of PCA is the human face recognition. Kirby and Sirovich [85] were the first to employ Karhunen Loeve transform to represent facial images. Their work was followed by the PCA Eigenface technique [164].

PCA (e.g. Eigenface method) [164] considers images as vectors in a high dimensional image space. All the images are projected onto the eigenspace spanned by the leading eigenvectors of the sample covariance matrix of the training images. Although PCA is popular in image feature extraction, it suffers from the following drawbacks: (i) 2D image matrices must be transformed into 1D image vectors. The resulting image vectors of faces usually lead to a high dimensional image vector space, where it is computationally intensive to compute the covariance matrix accurately due to its large size and the relatively small number of training samples. Other methods can be used to avoid covariance matrix computation (e.g. Adaptive methods as dis-

cussed in section 2.4), however the eigenvectors can be evaluated accurately by using covariance matrix only because the eigenvectors are statistically determined by the covariance matrix, irrespective of the method used for obtaining it [189], (ii) PCA does not make use of inherent matrix spatial structure of images, which may lead to low performance, (iii) Generalization ability is limited while extracting local features when variations in local region or a part of patterns are prominent, (iv) Because of the small sample size (SSS) problem, PCA is likely to be over-fitted to the training set.

To overcome the limitations of classical PCA for image data, more recently, Two-dimensional PCA (2DPCA) [189] (also known as image PCA (IMPCA) in the previous paper [187]) was proposed. 2DPCA was proved to be superior in terms of computational cost and classification. Unlike PCA that treats images as vectors, 2DPCA views an image as a matrix of image features. 2DPCA is more suitable for small sample size problems (like face recognition) since its image covariance matrix is quite small.

2DPCA description and algorithm: [189]

First, the covariance matrix, $(\mathbf{M})_{n\times n}$ is computed for the given set of training images, $\mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N\}$ as given by

$$(\mathbf{M})_{n\times n} = \frac{1}{N}.\sum_{i=1}^{N} E[(\mathbf{A}_i - \bar{\mathbf{A}})^T_{n\times m}.(\mathbf{A}_i - \bar{\mathbf{A}})_{m\times n}] \tag{2.1}$$

where $\bar{\mathbf{A}}$ is the mean of images and is given by $\bar{\mathbf{A}} = \frac{1}{N}.\sum_{i=1}^{N} \mathbf{A}_i$. Next, find $r\,(< n)$, eigenvectors of $\mathbf{M}$ corresponding to first r largest eigenvalues. Let $(\mathbf{E})_{n\times r}$ be the matrix of r column eigenvectors (projection vectors) chosen in this step. Finally, the set of images, $\mathbf{A}$, are projected onto $\mathbf{E}$ to get a set of reduced images, $\{\mathbf{IM}_i;$

$i = 1, 2, \ldots, N\}$ and is given by

$$(\mathbf{IM}_i)_{m\times r} = (\mathbf{A}_i)_{m\times n}.(\mathbf{E})_{n\times r} \tag{2.2}$$

Due to its computational superiority, 2DPCA is widely used in various fields especially in face recognition area. Junwei Tao et al [158] applied 2DPCA for palmprint recognition. Xiaoyu Zhang et al [197] used 2DPCA followed by Support Vector Machine (SVM) for face detection. Xiaoyu Zhang et al [197] showed that the method can effectively detect faces under complicated background, and the processing time is shorter than using SVM alone. Yin Hongtao et al [63] proposed a 2DPCA method using wavelets to obtain the feature vector representing a face: First, it uses the wavelet decomposition to extract intrinsic features of face images. As a result of wavelet decomposition, they obtained four sub-images (namely approximation, horizontal, vertical, and diagonal detailed images). It was shown that by decomposing a face image using wavelet transform, the low-frequency face image is less sensitive to the facial expression variations. The authors selected the approximation image as the feature of face image. Second, they performed 2DPCA twice (2DPCA in horizontal direction followed by vertical direction), after which the discriminant information is moved to the upper-left corner of the image and is used to classify the face image. The method showed significant reduction in computational time and slight improvement in recognition as compared to other wavelet methods [100][23].

Improving 2DPCA with different distance measures:

The typical classification measure used in 2DPCA-based face recognition is the sum of the Euclidean distance between two feature vectors in a feature matrix, called distance

measure (DM). However, this measure is not compatible with the high-dimensional geometry theory. So a new classification measure compatible with high-dimensional geometry theory and based on matrix volume is developed by Meng and Zhang for 2DPCA-based face recognition [109]. Distance between two images, $\mathbf{A}_i$, $\mathbf{A}_j$ using Volume measure is given by

$$\mathbf{VM}_{i,j} = \sqrt{det[(\mathbf{A}_i - \mathbf{A}_j)^T.(\mathbf{A}_i - \mathbf{A}_j)]} \tag{2.3}$$

Another method proposed by Zuo et al [201], combines 2DPCA with Assembled matrix distance (AMD) which was proved to be effective for 2DPCA based image recognition. Motivated by the idea of using matrix structure of 2DPCA, Chen et al [22] tried to extract features for any vector pattern by first 'matrixizing' it into a matrix pattern and then applying the matrixized version of PCA, known as MatPCA to the pattern. MatPCA uses a minimization of the reconstructed error for the training samples like PCA to obtain a set of projection vectors. It was observed that the computational burden of extracting features is largely reduced. However, it is to be noted that matrixizing a pattern may not work well with all the data sets because of its creation of artificial spatial relationships.

Some improvements to overcome large number of coefficients in 2DPCA:

Despite its superiority, 2DPCA needs more coefficients (that is more extracted features) for image representation than classical PCA. In an effort to reduce number of coefficients, Junwei Tao et al [158] applied classical PCA (1DPCA) to the coefficients (PCs) obtained by 2DPCA. In their method they selected the PCs that have better balance on maximizing the between-class distance while minimizing the within-

class distance rather than the principal components with highest variance, because such principal components are better for classification. Wen and Pengfei proposed an approach, IPCA [175] which uses 2DPCA to obtain the projective feature image which is processed by 2DPCA again. IPCA shows significant improvement in terms of recognition. Further in this direction of reducing number of coefficients, Zhang and Zhou proposed $(2D)^2$PCA, Two-directional 2DPCA, which computes projection vectors (eigenvectors) for both row and column directions [194]. The same bi-directional concept as shown by $(2D)^2$PCA [194] is re-iterated by different researchers: (i) Sun and Ruan [153] for face expression identification in the form of 2D-2DPCA, (ii) Konga et al [88] for recognition in the form of Bilateral-projection-based 2DPCA (B2DPCA). In the same paper Konga et al proposed a kernel version of 2DPCA called K2DPCA scheme and the relationship between K2DPCA and KPCA is explored and (iii) Anbang Xu in the form of complete PCA [181]. Zuo et al [202] extended assembled matrix distance (AMD) metric to improve Bidirectional PCA (BD-PCA) and an AMD metric is presented to calculate the distance between two feature matrices and then the nearest neighbour and nearest feature line classifiers are used for image recognition. Diagonal Principal Component Analysis (DiaPCA) [195] captures the essence of using the relationships between variations of rows and those of columns of images, using a different approach. In DiaPCA, for each training face image, a diagonal face image is computed as given in [195], then 2DPCA is applied for the set of diagonal face images.

Generalization to nDPCA:

Motivated by 2D forms of the PCA, Yu and Bennamoun further developed and ex-

tended the idea to an arbitrary n-dimensional space. Analogous to 1D- and 2DPCA, the new nD-PCA is applied directly to n-order tensors (n = 3) rather than 1-order tensors (1D vectors) and 2-order tensors (2D matrices). In order to avoid the difficulties faced by tensors computations, nD-PCA algorithm has to exploit a Higher-Order Singular Value Decomposition (HO-SVD) to make it practically feasible.

Discrimination front:

Although PCA ensures that the features extracted have least reconstruction error, it may not be optimal from a discrimination standpoint. To improve feature extraction discrimination point of view, (i) Nhat and Lee [111] proposed a 2DPCA based method which makes use of Laplacian weighting matrix method which considers data labeling, and makes the performance of recognition system better with the complexity nearly same as that of 2DPCA. Another direction to extract discriminant features from 2DPCA is to combine it with LDA technique. Sanguansat et al [142] combined 2DPCA and 2DLDA, which improved dimensionality reduction of the feature matrix in addition to improving classification accuracy. Similarly, Zuo et al [203] combined bidirectional PCA (BDPCA) with LDA (say, BDPCA + LDA), which performs an LDA in the BDPCA subspace. Their experimental results show that BDPCA + LDA needs less computational and memory requirements and has a higher recognition accuracy than PCA + LDA. In a different approach to improve 2DPCA further, Kim and Choi [81] computed window based 2DPCA and 2DLDA methods using image covariance obtained from windowed features of images. A windowed input feature consists of a number of pixels, and the dimension of input space is determined by the number

of windowed features. A $m \times n$ window feature is treated as a $m.n$ feature vector. Each element of an image covariance matrix can be obtained from the inner product of two windowed features. The 2DPCA and 2DLDA methods are then computed using the image covariance matrix of the windowed features. It was observed that 2DLDA performed well as compared to other LDA methods and 2DPCA. Using these window based 2D methods, (i) we can control the dimension of the input space by changing the window size or by overlapping the windows, which consequently solves the small sample size (SSS) problem and (ii) the computational load is significantly reduced.

Link to Block based (Feature Partitioning based) PCA approaches:

The methods based on 2D matrix structure (2DPCA) are proved to be equivalent [170] to special cases of image block based feature extraction (that is, when each row is taken as a block) (Section 2.2). Later Gao [183] proved that the earlier proof by Wang et al [170] is not correct. Gao analyzed that 2DPCA views the rows of images as training samples that constitute m sub-training sets instead of original images (m is equivalent to the rows of images) where as Block based PCA (modPCA) views entire $N.m$ blocks as a single training set.

Note that the approaches discussed in this section (except DWT based PCA) are based on whole patterns and have a common drawback of not exploiting local features. Local features are those extracted from a local region (or sub-pattern) rather than from entire image (or pattern).

2.4 Artificial Neural Network based Principal Component Analysis Methods

Batch methods versus Incremental methods:

Artificial Neural Networks (ANN) are massively parallel interconnections of simple neurons that function as a collective system. PCA computation can be done in two modes: (i) Batch mode (ii) Incremental or Adaptive mode. The batch methods assume that all the data is available beforehand. In batch methods the PCA is performed as follows: (i) Calculate covariance matrix by making use of the training data (patterns), (ii) the covariance matrix is then decomposed to find the principal component directions of the variances (that is eigenvectors corresponding to highest eigenvalues). In practice, usually the covariance matrix is diagonalized using some numerical technique such as Householder-QR technique [127]. In contrast to batch methods, Incremental methods (i) work directly with the data *without computation of covariance matrix* in advance and (ii) they might be implemented adaptively so that the directions of the Principal Components (PCs) are adjusted after a new data is received, without the need of reusing all previous data (patterns). This approach is suitable for real-time applications or for very high dimensional problems where the computational expense and storage requirement is an important consideration. One application area is computer vision, in which all visual filters are incrementally derived from very long on-line real-time video stream, motivated by the development of animal vision systems. On-line development of visual filters requires that the system perform while new sensory signals flow in. An online developing system must

observe an open number of images and the number is larger than the dimension of the observed vectors. There is evidence that biological neural networks use an incremental method to perform various learning, e.g., Hebbian learning [176]. As we all aware of, ANNs are well known for incremental learning. More details of Neural Networks for PCA can be found in [4] [37].

We recollect that PCA has two main properties - (i) It finds the uncorrelated directions of maximum variance in the data space and (ii) it provides the optimal linear projection in the least square sense. According to these two properties, two types of PCA networks can be found in the literature. Hebbian type learning algorithms are based on the variance maximization and uncorrelatedness property, whereas linear Auto-Associative MLPs (AA-MLP) compute the PCA space, because this subspace yields the best linear mean-square approximation [37].

Auto-Associative Neural Networks (AA-NNs) for PCA:

The relationship between PCA and AA-MLPs was first noticed by Bourlard and Kamp [13]. If AA-NN contains hidden layer size less than input layer size, the network works as feature extractor and finds efficient ways of compressing the information contained in the input patterns. The use of such a scheme for information compression and dimensionality reduction was first suggested by Rumelhart et al [139]. It was analyzed formally by Bourlard and Kamp [13] using the concept of singular value decomposition of matrices. Further results were obtained by Baldi and Hornik [4], who provided a complete description of the error surfaces of multilayer linear networks (of which AA-NNs with one hidden layer are a special case). Further PCA using AA-NNs extended to fuzzy data sets by Denoeux and Masson [34].

This method exploits recent results regarding the ability of linear AA-NNs to perform information compression in just the same way as PCA, without explicit matrix diagonalization. Further, Girard and Iovleff [51] proposed auto-associative models to generalize PCA. These AA-NN models have been introduced in data analysis from a geometrical point of view. They are based on the approximation of the observations scatter-plot by a differentiable manifold. In their study those models are interpreted as projection pursuit models adapted to the auto-associative case. The supervised AA-NN algorithms, may end up being trapped into local minima [13], and also their global treatment of information makes it difficult to implement ANNs into efficient hardware [60]. Therefore many researches focussed on the study of unsupervised ANNs, particularly Hebbian type ANNs (after the work of the Canadian neurophysiologist Hebb) [79][116][117][149][141][177]. These methods are based on Oja's earlier work [114].

Hebbian-Type ANNs for PCA (HANN):

The motivation for the popular Hebbian ANNs came from the so called Hebbian Learning Rule. In his seminal work 'The Organization of Behavior' [59], Hebb proposed a simple, yet biologically motivated rule, for adjusting the synaptic weights during a neural network learning process, which is given as *'when neuron N_1 and unit N_2 are simultaneously excited, increase the strength of the connection between them'*. For the case where neurons are modeled as units with continuous output activation this correlation-type rule is given in mathematical form as *'Adjust the strength of the connection between units A and B in proportion to the product of their simultaneous activation'*. Interestingly, this simple rule turns out to be closely related to PCA when

the neural units are linearly modeled. Oja [114] showed that a normalized version of the Hebbian rule applied on a single linear unit extracts the first principal component of the input sequence, i.e., it converges to the principal eigenvector of the input auto-correlation matrix [92] [112]. Recently, Nicole [112] evaluated Subspace Network (SN) [117], Generalized Hebbian Algorithm (GHA) [117][141], Weighted Subspace Algorithm (WSA) [117] and Stochastic Gradient Ascent (SGA) [117] ANNs, in terms of efficiency of extraction of eigenvectors and pattern classification, compression and reported the following facts. It was observed that there is a decrement in the accuracy of computation of the first eigenvector along the number of neurons for all the ANNs, with the relevant exception of WSA. For classification tasks, Hebbian ANNs are unable to distinguish patterns properly as compared to SVD based PCA. It is worth noting the performance of the SN algorithm in terms of compression and reconstruction, when only very few (actually, 4) eigenvectors are considered, is quite comparable to the SVD result. The other ANNs perform grossly worse; their NMSEs (normalized MSE) are by and large an order of magnitude larger than that from SVD. As the number of eigenvectors considered increases, though, the gap between the SVD algorithm's performance and the ANNs' grows and the precision of the reconstruction by the non-adaptive algorithm's results increases. In a nutshell, the results obtained for more demanding tasks suggest that the ANNs (and also WSA) are easily outperformed by other classical numerical algorithms, especially whenever a high precision in the reconstruction is requested.

Kung et al [92][91] proposed the Adaptive Principal-component Extractor (APEX) model which can effectively support a recursive approach for the calculation of the

p^{th} principal component given the first $(p-1)$ ones. The motivation behind such an approach is the need to extract the principal components (PCs) of a given data patterns when the number of required PCs is not known a priory. It is also useful in applications such as speech analysis where the correlation matrix (covariance matrix) of the data might be slowly changing with time. Then the new PC may be added to compensate the change without affecting the previously computed PCs. APEX model has both feed-forward and lateral connections.

The methods based on Oja's original work [114] may not adequately consider automatic selection of learning parameters, thus leading to slow convergence or even divergence if parameters are not properly chosen. The stability and the ways of choosing the values of the learning rate parameters of the Oja's one-unit learning rule and some other gradient type algorithms have been discussed in [118] [78] [31]. Chen and Chang [19] proposed an adaptive learning algorithm (ALA) for PCA, where in, the learning rate parameters can be selected automatically and adaptively according to the eigenvalues of the input covariance matrix that are estimated during the learning process. The simulation results demonstrated that the ALA can converge quickly to the desired targets while the GHA diverges in the large eigenvalue case. Further, based on the work of Oja [118] and Sanger [141], Weng et al [176] proposed a fast converging method, candid covariance-free incremental PCA, to compute the principal components of a sequence of samples incrementally without estimating the covariance. The method is motivated by the concept of statistical efficiency (the estimate has the smallest variance given the observed data). To do this, it keeps the scale of observations and computes the mean of observations incrementally, which is

an efficient estimate for some known distributions such as Gaussian. The method is proposed for real-time applications, and does not allow iterations. It converges very fast for high dimensional image vectors. Chatterjee et al [17] presented adaptive algorithms which are based on an unconstrained objective function, which can be minimized to obtain the principal components. By using this objective function, they derived adaptive algorithms by using: (i) gradient descent, (ii) steepest descent, (iii) conjugate direction and (iv) Newton-Raphson methods for PCA. These methods were shown to converge faster than the traditional gradient descent PCA algorithms due to Oja, Sanger, and Xu [17].

In local Hebbian type learning algorithms the modification of the i^{th} row of the weight matrix between input and output layer depends only on the i^{th} output unit and the input. Due to this locality it has been argued that these algorithms are biologically plausible [174].

The detailed discussion of generalized Neural Network PCA models such as constrained PCA, oriented PCA, asymmetric PCA and other ANN PCA models can be found in [36].

Nonlnear PCA using Neural Networks:

Because of its linearity, PCA is not always suitable, and has redundancy in expressing data. To overcome this problem, some nonlinear PCA methods have been proposed. Whether the nonlinear approach has a significant advantage over the linear approach is highly dependent on the data set. The nonlinear approach is generally not good if the data is short and noisy, or the underlying structure is essentially linear. Nonlinear principal component analysis (NLPCA) using AA-NNs was first introduced by

Kramer [90] in the chemical engineering literature, and is now used by researchers in many fields. The presence of local minima in the cost function renders the NLPCA using ANNs somewhat unstable, as optimizations started from different initial parameters often converge to different minima. Regularization by adding weight penalty terms to the cost function is shown to improve the stability of the NLPCA [65]. Another way to realize non linearity is to have a mixture model [193] by concurrently performing global data partition and local linear PCA. The partition is optimal or near optimal, which is realized by a soft competition algorithm called 'neural gas'. The local PCA type representation is approximated by a neural learning algorithm in a nonlinear auto-encoder network, which is set up on the generalization of the least-squares reconstruction problem leading to the standard PCA. Such a local PCA type representation has a number of numerical advantages, for example, faster convergence and insensitive to local minima. Most of the nonlinear methods have drawbacks, such that the number of principal components must be predetermined, and also the order of the generated principal components is not explicitly given. Ryo Saegusa et al [140] proposed a nonlinear PCA algorithm based on hierarchical MLP neural network model that nonlinearly transforms data into principal components, and at the same time, preserving the order of the principal components. The network composed of a number of independent sub-networks that can extract ordered nonlinear principal components.

Other methods:

A comparative study of derived classification accuracies of a neural network (NN) implementation of Sammon's mapping, an auto-associative NN (AA-NN) and a mul-

tilayer perceptron (MLP) feature extractor and conventional principal component analysis (PCA) was carried out [99]. The study reveals that MLP provides the highest classification accuracy at the cost of deforming the data structure, whereas the linear models preserve the structure but usually with inferior accuracy. Huang [66] used Generalized Hebbian algorithm (GHA) to perform PCA for *Seismic data analysis*. The neural network using an unsupervised GHA is adopted to find the principal eigenvectors of a covariance matrix in different kinds of seismograms. GHA can extract the information of seismic reflection layers and uniform neighbouring traces. The method also provides a significant seismic data compression [66].

From our study, it is understood that, incremental learning based on ANNs for PCA is very useful for applications where the data is received incrementally. These methods also reduce storage requirements and computational requirements as well. However, at times incremental learning methods may suffer from the drawbacks: (i) slow convergence (in this case these methods may be computationally expensive), (ii) not efficient to extract eigenvectors, (iii) may not be efficient for compression and reconstruction and (iv) may not be good for pattern classification view point as compared to classical PCA methods. Batch methods although require relatively more storage and computational requirements, they show their efficiency in applications (where data is available beforehand) in terms of (i) better extraction of eigenvectors, (ii) better feature extraction and construction and (iii) pattern recognition tasks.

2.5 Kernel Principal Component Analysis Methods

The kernel principal component analysis (KPCA) has been applied in numerous machine learning applications and it has exhibited superior performance over previous approaches, such as PCA. Classical PCA is suitable in applications where the underlying structure is linear. The linear PCA either needs more principal components or unsuitable for the data sets where nonlinear structure is present. KPCA [145] introduces a nonlinear form of doing PCA by using *kernel functions*. In KPCA, as a first step we map the input feature space into a kernel space by using a kernel nonlinear mapping, next we perform classical PCA on the transformed kernel space.

$$\phi : \Re^d \rightarrow \mathbf{K} \tag{2.4}$$

where $\mathbf{X}_i \in \Re^d$ is input feature vector and $\mathbf{K}$ is high-dimensional transformed kernel space. In fact, we do not actually compute $\phi - map$ for an input data $\mathbf{X}_i$, instead we use kernel functions in the place of dot products $\phi(\mathbf{X}_i).\phi(\mathbf{X}_j)$. Some kernel functions include (i) polynomial kernel which is given by

$$k(\mathbf{X}_i, \mathbf{X}_j) = (\mathbf{X}_i.\mathbf{X}_j)^n \tag{2.5}$$

(ii) radial basis functions (iii) sigmoid kernels, etc. Compared to other nonlinear methods to PCA, for e.g. Auto Associative MLPs or principal curves, KPCA has the advantage of not needing nonlinear optimization, it needs only solving eigenvalue problem as classical PCA. Thus there is no problem of getting trapped into local minima during learning in case of KPCA. KPCA as a nonlinear feature extractor has

been proved to be a powerful tool in preprocessing for classification tasks. Mika et al [110] tried to emulate KPCA as a natural generalization of linear PCA. They have shown how to use nonlinear features for data compression, reconstruction, and denoising applications common in linear PCA. Please note that as the results provided by KPCA live in some high dimensional feature space and need not have pre-images in input space. Their experiments reveal that reconstruction results of KPCA were comparable with linear PCA and KPCA has shown significantly better results with respect to denoising. More details including derivation on KPCA can be found in [146].

Some improvements to KPCA:

KPCA has high computational cost due to its dense expansions of kernel functions. To overcome this shortcoming, Sparse Kernel Feature Analysis (SKFA) method was proposed by Smola et al [150]. SKFA overcomes the problem by using L_1 norm for feature extraction in coefficient space, instead of kernel Hilbert space in which KPCA is formulated in. The SKFA algorithm was proved to be fast and leads to sparse representations. KPCA is not sparse because the computation of principal components require to compute kernel functions associated with every training patterns. Another approach to overcome this problem is proposed by Tipping [162] which is known as Sparse Kernel Principal Component Analysis (SKPCA). First the SKPCA method approximates covariance matrix in feature space by computing a subset of outer products of feature vectors using maximum likelihood approach based on probabilistic PCA [163]. Next, the KPCA is applied to obtain sparse projections.

The most widely used kernel functions in the literature are polynomial kernels,

Gaussian kernels, and sigmoid kernels. In an effort to improve face recognition performance Chengjun Liu proposed Gabor-based KPCA with new kernel function 'Fractional Power Polynomial Models' [102]. The method integrates Gabor wavelet representation of face images with KPCA using fractional power polynomial models for enhanced face recognition. The feasibility of the Gabor-based KPCA method with fractional power polynomial models has been successfully tested on both frontal and pose-angled face recognition, using two data sets from the FERET database and the CMU PIE database. The superiority of the Gabor-based KPCA method with fractional power polynomial models is shown by the author in terms of both absolute performance indices and comparative performance against PCA, KPCA with polynomial kernels and KPCA with fractional power polynomial models, etc. To avoid the potential problems with standard KPCA, Chin and Suter [24] brought out an incremental computation algorithm for KPCA, where incremental linear PCA is computed in the kernel induced feature space.

Choosing number of dimensions in Kernel based Subspace methods:

One of the traditional approaches to select the dimensions is based on cumulative proportion computed from the kernel matrix for each class. To select number of dimensions systematically, Kim et al [84] proposed a new method selecting optimal or near-optimal subspace dimensions for KNS classifiers using a search strategy and a heuristic function called the 'Overlapping Criterion'. The heuristic criterion that uses critical information about the specific dimensions chosen, called the overlap, between the corresponding subspaces.

Subspace classification using kernel trick:

Peng Zhang et al [196] proposed a kernel-pooled local discriminant subspace method and compared it against KPCA and generalized discriminant analysis (GDA) in classification problems. The method computes a nonlinear pooled local discriminant subspace by using the kernel trick and it makes use of Gaussian kernel.

Some applications of KPCA:

KPCA with polynomial kernel of degree d, is applied to face recognition by Yang et al [190] and it was observed that KPCA with kernel of degree 3 has given a lower error rate in both Yale (Eigenface: 28.49%; KPCA ($d = 3$) : 24.24%)) and AT&T (Eigenface: 2.75% KPCA ($d = 3$) : 2%) data sets as compared to traditional Eigenface method. They carried out experiments using leave-one-out strategy. KPCA is also used by Kim et al [83] for face recognition application using ORL face data set. It was showed that KPCA with polynomial kernel of degree 4 (error rate is : 2.5%) is significantly lower error rate than traditional PCA (error rate: 10%). In natural language domain, KPCA is used by Dekai Wu et al [178] for word sense disambiguation. The method outperformed other methods such as naive Bayes method, maximum entropy method and SVM model as well. KPCA is also used in remote sensing domain [156] and scene analysis for mobile robot based on multi sonar ranger data [171].

2.6 EM Algorithms for PCA

Expectation Maximization (EM) algorithm can be used for learning the principal components of a data set. EM algorithms do not require computing the sample covariance matrix and deal with high dimensional data more efficiently than classical PCA. Tipping and Bishop [163] proposed probability model for PCA. PCA can be

viewed as a limiting case of a particular class of linear Gaussian models. Gaussian models assume that $\mathbf{x}$ is produced as a linear transformation of some r-dimensional latent variable $\mathbf{y}$ plus additive Gaussian noise, $\mathbf{v}$. Based on the probability model, EM algorithm is used to learn principal component directions [138]. An application of Probabilistic PCA for rapid speaker adaptation may be found in [82]. Other EM algorithms include, EM algorithm for integrated-squared-error minimization [1], EM algorithm for high-dimensional spaces [38]. The EM learning algorithm for PCA is an iterative procedure for finding the subspace spanned by the leading r eigenvectors without explicit computation of the sample covariance. It is attractive for small values of r, because its complexity is limited by $O(r.N.d)$ per iteration and depends only linearly on both the dimensionality of the data (d) and the number of points (N). Methods that explicitly compute the sample covariance matrix have complexities limited by $O(N.d^2)$.

2.7 Hybrid Methods

In this section we review the techniques which are combination of PCA and other methods such as LDA, Rough Sets theory.

PCA and LDA:

Zhao et al applied LDA for principal components obtained from PCA technique to improve the generalization ability of LDA when few samples are available. The combined PCA+LDA linear classifier shows significant improvement over pure LDA linear classifier [199]. It is well-known that LDA does not work well with nonlinear applications. To improve LDA for nonlinear case, Yang et al [186] [185] investigated Kernel

Fisher Discriminant (FLD) and then proposed two-step method: first KPCA is applied and then LDA is applied in the KPCA-transformed space. Rajagopalan et al [130] proposed a face recognition method that combines information acquired from global and local features of the face for improving performance. They have considered the 3 complementary features, (i) grayscale image of the face, (ii) the edginess image of the face (which is robust to different illuminations), and (iii) the eyes (which are robust against occlusions and facial expressions). Dimensionality of each of 3 original feature spaces is reduced by applying PCA followed by fisher analysis. Recognition is done by probabilistically fusing the confidence weights derived from each feature space. This method was proved to be quite good as compared to other methods which use only single feature (i.e gray-scale image only or eyes or edginess image only). However, the problem with this approach is computational overhead because it needs to compute 3 different feature spaces. Another interesting combined classifier (called MPL) is proposed by Alok Sharma et al [147]. MPL classifier is a combination of Minimum Distance (MDL), class-dependent PCA and LDA methods.

PCA and Rough Sets Theory:

Pawlak first advocated the rough set theory (RS) as an approach to automatic knowledge acquisition in 1982 [121]. Using RS theory one can find attribute dependencies in database-like information systems, for e.g. a decision table. The basic idea is to compute decision or classification rules through data attribute and attribute-value reductions. RS constrains the data reductions by keeping the discernibility relations among data objects in the table unchanged. Swiniarski and Skowron [154] proposed a method for feature selection that uses Rough Set (RS) theory to the PCA result.

The approach projects the original d-dimensional patterns data, $\mathbf{X}$ into the reduced r-dimensional patterns ($r < d$), $\mathbf{Y}$ in the principal component space. It then makes the reduced and projected data set with real valued attributes discrete and then computes a attribute reduct which comprises of projected features. It is known that, there exists cause-effect relationship between condition and decision attributes in a decision table. The attribute reduct that represents the condition attributes with a larger contribution to cause better than the attribute reduct with smaller contribution. The rule set derived from attribute set with maximal contribution is expected to be more resistant to noise and stronger in its generalization capabilities. To find the contribution of attributes, PCA is one of the well known methods in the pattern recognition literature. Zeng et al [192] proposed an approach KA-RSPCA which is based on PCA in combination with Rough Sets Theory. They used PCA to rank importance of condition attribute by using Cumulative Correlation Coefficient (CCC).

2.8 Methods to Choose Number of Principal Components

PCA is able to retain meaningful information in the early axes whereas variation associated to experimental error, measurement inaccuracy, and/or rounding is summarized in later axes [50]. PCA is able to identify relationships by computing principal components (which are linear combinations of variables) showing common trends of variation can contribute substantially to the recognition or classification of patterns in the data. However, the issue of determining whether or not a given

axis (i.e. principal component) summarizes meaningful variation is not clear in many cases. Please note that when the correct number of non-trivial principal components is not retained for subsequent analysis, either relevant information is lost or noise is included, causing a distortion in underlying patterns of variation/covariation [41][95]. Determining the number of non-trivial principal components remains one of the greatest challenges in providing a meaningful interpretation of multivariate data and has been a long-standing issue in both biological and statistical literature [76].

Methods based on confidence intervals:

Parallel Analysis (PA)[44] involves a Monte Carlo approach to generate a large number of eigenvalues based on simulated data sets that are equivalent in size to the observed data set of interest, and comprises of independent normally distributed variables. These eigenvalues are then used to build confidence intervals for each axis. If observed values exceed the critical value, then we reject the null hypothesis according to the pre-specified significance level. Parallel analysis is based on independent normally distributed data and is parametric. Other methods which are distribution-free include randomization and bootstrap that may give a more robust assessment for non-normal distributions.

Randomization methods based on eigenvalues follow the protocol: (i) randomize the values within variables in the data matrix, (ii) conduct a PCA on the reshuffled data matrix, and (iii) repeat steps (i) and (ii) a total of 999 times. In each randomization, we can evaluate test statistics based on the eigenvalues such as (i) the observed eigenvalue for an axis [159], (ii) a Pseudo-F-ratio is calculated as each eigenvalue divided by the sum of the remaining (or smaller) eigenvalues [160]. Similarly one can

have *randomization methods based on eigenvectors.*

Bootstrap methods based on eigenvalues. Bootstrap confidence intervals for eigenvalues [70] are computed based on resampling the original data with replacement so that the bootstrapped sample is consistent with the original dimensions of the data matrix. 1000 bootstrapped samples are drawn and a PCA is conducted on each of them. There are a number of methods for estimating confidence intervals and one of them is the percentile method [105]. In the similar way, *bootstrap methods based on eigenvectors* compute bootstrap confidence intervals for loadings instead of eigenvalues [122].

Correlation critical values for eigenvectors method tests loadings against the critical values for parametric correlation from standard statistical tables. Any particular axis with at least two significant eigenvector loadings is retained. Other methods include Bartlett‘s test for the first principal component [5], Lawley‘s test for the second principal component [94]. Recently Chen [20] proposed a confidence interval using a stepwise selection procedure for the number of important principal components in PCA. An i^{th} principal component important if λ_i/λ_1 is closer to 1, where λ_1 is the largest eigenvalue.

Methods based on average test statistic values:

Rules based on average values assess whether an observed test statistic based on eigenvalues or eigenvectors is larger than the average value expected under the null hypothesis of no association between variables. According to *Kaiser-Guttman method* [54], when using correlation matrices, population components having eigenvalues larger than 1.0 should be retained. In *Broken-stick* method one assumes that the total

variance in a multivariate data set is divided at random amongst all components, the expected distribution of the eigenvalues can be assumed to follow a broken-stick distribution [98]. The idea underlying the model is that if a stick is randomly broken into p pieces, b_1 would be the average size of the largest piece in each set of broken sticks, b_2 would be the average size of the second largest piece, and so on. If the k^{th} component has an eigenvalue larger than b_k, then the component is retained. *Random average under permutation* method is based on the average eigenvalue obtained under a randomization of the data matrix. If the observed value exceeds the average random value, that particular axis is to be retained. Peres-Neto et al [123] conducted a comparative study 20 stopping rules (to find number of PCs) and found that (i) Random average under parallel analysis, (ii) Random average under permutation, (ii) Parallel Analysis, (iii) $Rnd-Lambda$, (iv) $Rnd-F$ or (v) Minimum Average partial correlation [167] methods perform well as compared to other methods. They also proposed a two-step approach: First, a Bartlett's test is used to test the significance of the first principal component, indicating whether or not at least two variables share common variation in the entire data set. If first PC is significant, a number of different rules can be applied to estimate the number of non-trivial components to be retained. However, the relative merits of these methods depend on whether data contain strongly correlated or uncorrelated variables.

Other methods:

The interpretation of the principal components is generally based on the respective magnitudes of the loadings assigned to the variables. The simplification of the principal components is a great concern for experts in many areas. Vines [169] proposed

a procedure that achieves a simplification of the principal components by seeking approximate components that can be represented by integers. Vigneau and Qannari [168] discussed a strategy of simplification based on cluster analysis of variables. On the similar lines to Jeffers [74], Ledauphin et al [96] proposed a method of hypothesis testing to ascertain the significance of principal components and the variable contributions to the determination of the principal components. If a variable contribution turns out to be non-significant then the loading associated with this variable is set to zero. This process simplifies interpretation of principal components. Hypothesis testing is based on a procedure of simulation by permutations of the rows (each row corresponds to a variable) of the data matrix at hand.

Cadima et al [14] discussed computational aspects of several algorithms for the optimization problems resulting from three different criteria (RM, RV and GCD criteria) in the context of PCA. They found that the local search methods (e.g. local improvement, simulated annealing and genetic algorithms) performed significantly better than the greedy-type algorithms (e.g. forward selection, backward elimination and stepwise algorithms with a default forward or backward direction).

2.9 Comparison of PCA with Other Feature Extraction Methods

PCA versus LDA:

In the context of the appearance-based paradigm for object recognition, it is generally understood that LDA based algorithms are superior to those of PCA based algorithms

because LDA deals with class discrimination. However, empirical evidence by Martonez and Kak [108] shows that this is not always the case and PCA can outperform LDA when the training data set per class is small. It is also showed that PCA is less sensitive to different training data sets. Beveridge et al [10] compared Nearest Neighbor classifiers using principal component and linear discriminant subspaces using different choices of distance metric. They computed probability distributions for algorithm recognition rates and pairwise differences in recognition rates using a permutation methodology. They found that the principal component subspace with Mahalanobis distance performed well and next better performance is using $L2$ distance metric. Linear discriminant subspace is less sensitive to the choice of distance metric, and its performance is always worse than the principal components classifier using either Mahalanobis or L1 distance. Probability distributions for recognition rates and differences in recognition rates relative to different choices of gallery and probe images have been created using a Monte Carlo sampling method. Other comparisons between PCA and LDA may be found in [7] [9].

PCA versus ICA:

Yang et al [188] investigated the two architectures of ICA for image representation and found that ICA Architecture-I involves a PCA process by vertically centering (PCA-I), while ICA Architecture-II involves a whitened PCA process by horizontally centering (PCA-II). These two PCA versions are used as baseline algorithms to evaluate the ICA-based face recognition systems. They found through their experimentation on FERET face data that there is no significant performance differences between ICA Architecture-I (II) and PCA-I (II), and ICA Architecture-II significantly

outperforms the standard PCA. Also the recognition performance of ICA, whether using Architecture-I or II, strongly depends on its involved PCA process (PCA-I or II). The pure ICA projection seems to have little effect on the performance of face recognition. However, Baek et al [2] have tested three different distance metrics L1 norm, L2 norm, and cosine angle - for both PCA and ICA. Baek et al found, contrary to previous reports in the literature, that PCA significantly outperforms ICA when the best performing distance metric (L1 norm in this case) is used for each method. In another development, Fortuna et al [42] compared PCA, ICA, KPCA and FLD with respect to accuracy of visual position measurement and they examined to see the ability of the methods to discriminate positions in a 2D visual subspace. The comparison is done both constant and varying illumination and random occlusion. It was shown that PCA provides overall good performance compared with more sophisticated techniques such as ICA, FLD, and KPCA at a reduced computational complexity. Connie et al [27] performed comparison of PCA, ICA, LDA and Wavelet method on palmprint data and found that application of LDA on wavelet sub-band is able to yield low FAR and FRR rates.

2.10 Some Applications of PCA

Face tracking and recognition:

Turk and Pentland [164] presented an eigenfaces approach to develop a near-real-time face recognition system which tracks subject's head and recognizes the person by comparing the face of the person with those known individuals. Face images are projected onto face space, that best encodes the variation among the faces. The face

space is given by eigenvectors of the training set of faces. A statistical assessment of subject factors was done by Givens et al [52] in PCA recognition of human faces. This study considered 11 factors that might make recognition easy or difficult for 1072 human subjects in the FERET dataset. The factors considered include race (white, Asian, African-American, or other), gender, age (young or old), glasses (present or absent), facial hair (present or absent), etc,. An ANOVA is used to determine the relationship between these subject covariates and the distance between pairs of images of the same subject in a standard *eigenfaces subspace*. Some outcomes of their study include (i) the distance between pairs of images for subjects decreases for people who consistently wear glasses, so wearing glasses makes subjects more recognizable, (ii) Pair-wise distance also decreases for people who are either Asian or African-American rather than white.

Astronomical applications:

Discrimination of Giant and Dwarf Spectra in K-stars. Ibata et al [67], used a variant of PCA for discrimination problems in astronomy. They have presented the problem of discrimination between K-giant and K-dwarf stars from intermediate resolution spectra near the Mg 'b' feature. For the highest S/N spectra, the automated classification agrees very well (at the $90-95\%$ level) with the visual classification.

PCA of the Lick indices of galactic globular clusters. Strader and Brodie [152] applied PCA of high-quality Lick/IDS absorption-line measurements for 11 indices in the wavelength range $4100-5400A$ for 39 galactic globular clusters (GCs). Only the first principal component appears to be physically significant. It was found that there is a tight linear relationship between this first component (PC1) and GC metallicity

over a wide range in $[m/H]$ ($-1.8 \leq [m/H] \leq 0$), suggesting that PC1 can be used to accurately estimate metallicities for old extra galactic GCs from their integrated spectra. It was found that little evidence for substantial differences in broad abundance patterns among galactic GCs.

PCA of speech spectrogram images:

The sound spectrogram is a commonly used three-dimensional (time-frequency-intensity) representation of an acoustic signal. Fourier descriptors (FDs) have been proved very useful for characterizing the boundary of segmented isolated words containing the English semi-vowels /w/, /y/, /l/, and /r/. Pinkowski [125] investigated the relevance of 16 32-point FDs combined with 17 other general features (to characterize non-shape parameters) for classifying objects contained in binary spectrogram images. PCA is used for reducing dimensions on a speaker-dependent data set consisting of 80 sounds representing 20 speaker-dependent words containing English semi-vowels. With only eight features, including four 32-point FDs and four general features obtained from PCA, a 97.5% recognition rate is obtained. The appropriateness of shape descriptors alone for classifying spectrogram objects may be enhanced if they are combined with other features, particularly those containing information on orientation (principal axes). Orientation features can be obtained from the eigenvector and chain code representations on binary objects.

Pattern classification using PCA and fuzzy rule bases:

Ravi et al [134] have used PCA to get principal components, which are subsequently fed in fuzzy rule based classifier as new set of features. This process has given a very high classification rate of 100% in leave-one-out technique with a few rules in the case

of some aggregators. The chosen principal components accounted for only 89% and 92% of the total variance in Wine and Breast cancer data sets respectively. Using this study as an evidence, PCA can be used as a useful alternative to other methods of feature selection existing in the literature while solving classification problems of higher dimensions using fuzzy rule based classifiers.

PCA-based branch and bound search algorithms for computing k nearest neighbours: Searching the k nearest neighbors in a multi-dimensional vector space is a very common phenomenon in pattern recognition. Recently, several branch and bound search algorithms were proposed that use a decomposition method based on PCA. These algorithms search the nearest neighbors in a vector space where the dissimilarity between two vectors is expressed by the Euclidean distance. It was shown that these algorithms have a linear space complexity, an average number of distance computations bounded by a constant term and a time complexity that is very close to logarithmic for a small number of dimensions. The most important aspects that influence the efficiency of the search algorithm are: (i) the decomposition method, (ii) the elimination rule, (iii) the traversal order and (iv) the level of decomposition. A theoretical derivation of an efficient decomposition method based on PCA is given by Dohaes et al [55]. Then, different elimination rules and traversal orders are combined resulting in different search algorithms.

PCA to facilitate fast detection of transient-evoked otoacoustic emissions: Transient-evoked otoacoustic emissions (TEOAE) are acoustic signals coming from the inner ear (outer hair cells of the cochlea) after acoustic stimulation by clicks and tone-bursts. These responses can be recorded from the ear canal of all normal adults,

children, and neonates shortly after birth, and are used as a clinical test to assess the integrity of the peripheral organ in TEOAE based newborn hearing screening programs. Some of their potential applications (e.g., their use as a tool in newborn hearing screening programs) are deeply related to the duration of each recording session. This duration can be strongly reduced by applying PCA approach to a set of TEOAE recorded from the same ear at different stimulus levels averaging only a few sweeps (a maximum of 100 versus the classical 260). The PCA approach used here is able to enhance the signal-to-noise ratio and, in turn, to allow a correct detection of the responses. The application of the PCA approach to a set of TEOAE recorded at different stimulus levels reduces on average the acquisition time of TEOAE to about one fourth of the time with the classical procedure. The comparison between the Similitude values provided statistical evidence that the PCA approach produces no loss of information in the set of data in terms of similarity between the rapidly acquired PCA-processed set and the GS set. The use of the PCA approach statistically improves the reproducibility of the set of data both for 60- and 100-sweep averaged data and the PCA approach improves dramatically the identification of the response in these conditions [133].

Machine Defect Classification:

Sensor-based machine condition monitoring has gained increasing attention from the research community world-wide. The goal of machine condition monitoring is to obtain operational status of the machines and use the information to (i) identify potential machine faults and failure before they occur, thus reducing unexpected and costly machine downtime, and (ii) better control the quality of products, which is

closely related to the condition of the machine. The information gathered from the monitoring sensors ultimately provide insight into the manufacturing process itself, enabling effective high-level decision-making for quality production at a lower cost. The PCA-based feature selection scheme for machine condition monitoring is based on the understanding that the amplitude of vibration signals of defective machine components increases as the severity of the defect increases. The issue of feature selection from a contending feature set arises, because of the stochastic nature of the defect propagation in machinery. Generally, as the defect severity increases, an overall increasing vibration trend is superimposed by local variations of smaller magnitudes. The goal of feature selection is therefore to select features that allow for an accurate description of the defect condition, and subsequently, reliable defect classification, diagnosis, and prognosis. The PCA approach was developed to reduce the dimensionality of the input features for both supervised and unsupervised classification purposes [104].

PCA to improve fault detection and classification (FDC) performance:

Yue et al [191] proposed sample-wise weighted PCA and variable-wise weighted PCA. Sample-wise weighted PCA is used to address issues with model updating. By adapting models with process changes, the long-term validity of a PCA model can be maintained. Variable-wise weighted PCA is used to incorporate process and sensor knowledge. PCA models built this way require less maintenance and result in better FDC performance. Yue et al [191] performed comparison studies on plasma etcher FDC and had shown that weighted PCA can adapt to process drift, reduce the occurrence of false alarms and make the models easy to maintain.

Computer-aided drug design:

Molecular similarity, as an important tool of computer-aided drug design has developed rapidly. Its calculation has also been developed from planar, rigid, 2D molecules to steric, flexible, 3D molecules. However, 3D molecular similarity calculation is easy to fall into local optima and the calculation is always time-consuming. Xian et al [179] proposed a method of flexible 3D molecular similarity calculation through the evaluation of molecular electrostatic potentials (MEP) with PCA, genetic algorithm (GA) and tabu search (TS). PCA is used to preprocess, GA is used to align two molecules and TS is used to decrease the probability of falling into local optima. The authors calculated the molecular similarities of benzene and its derivatives, a group of insecticides and a series of acetylcholinesterase inhibitors.

Discrimination of varieties of tea using near infrared spectroscopy by PCA and BP model:

Visible/near-infrared (Vis/NIR) spectroscopy , with the characteristics of high speed, non-destructiveness, high precision and reliable detection data, etc., is a pollution-free, rapid, quantitative and qualitative analysis method. A new approach for discrimination of varieties of tea by means of Vis/NIR spectroscopy $(325-1075nm)$ was developed by Yong He et al [58]. In this approach, the spectral data is compressed by the wavelet transform (WT). The features from WT can be visualized in principal component (PC) space, which can lead to discovery of structures correlative with the different class of spectra samples. It appears to provide a reasonable clustering of the varieties of tea. The scores of the first eight principal components computed by PCA had been applied as inputs to a back propagation neural network with one hidden

layer. The 200 samples of eight varieties were selected randomly to build BP-ANN model. This model is used to predict the varieties of 40 unknown samples. The recognition rate of 100% is achieved.

PCA-based web page watermarking:

The tamper-proof of web pages is of great importance. Some watermarking schemes have been reported to solve this problem. However these watermarking schemes and the traditional hash methods have a problem of increasing file size. Zhao and Lu [198] proposed a novel watermarking scheme for the tamper-proof of web pages, which is free of this embarrassment. For a web page, the proposed scheme generates watermarks based on PCA technique. PCA is applied on a matrix produced from a web page and a secret key. The watermarks are then embedded into the web page through the upper and lower cases of letters in HTML tags. When a watermarked web page is tampered, the extracted watermarks can detect the modifications to the web page, thus we can keep the tampered one from being published.

Remote Sensing Applications:

Aerosols are liquid and solid particles suspended in the air from natural or man-made sources. They can affect human health, visibility, and climate. Aerosol particles affect climate directly by reflecting and absorbing solar and terrestrial radiation, and indirectly by their influence on cloud micro-physics. Zubko et al [200] applied PCA to estimate how much information about atmospheric aerosols could be retrieved from solar-reflected radiance observed over oceans by a satellite sensor as a function of the number of wavelength bands, viewing angles, and stokes parameters. The following quantities are used to vary: aerosol optical thickness, single-scattering albedo (SSA)

of aerosol particles, height of the aerosol layer, aerosol model (includes size distribution parameters and optical properties), and wind speed. The real refractive index is kept constant and, therefore, is not part of the analysis. To calculate the number of significant principal components (PCs), the cumulative percent variance rule is used, which takes into account anticipated errors of measurements. The reported results predict how much additional information can be retrieved from observations by adding more wavelength, angle, and polarization channels. For example, for the moderate resolution Imaging Spectro-Radiometer instruments, the number of significant PCs is 2 to 3; for multi-angle Imaging Spectro-Radiometer, 3 to 5, etc. The calculations show that the observations should be most sensitive to the aerosol model followed in decreasing order by optical thickness, SSA, and aerosol height. It is found that there is no systematic increase in the information about aerosol starting from $10 - 15$ view angles for unpolarized observations and 30 view angles for those with linear polarization. It is achievable with modern detectors to retrieve up to 10 and 16 significant PCs from unpolarized and polarized observations respectively. The methodology and results of PCA can be useful for estimating the reliability of aerosol parameters retrieved from existing and future satellite observations.

A discussion of PCA in remote sensing may be found in [6].

2.11 How do the Existing PCA Methods Address the Problems of Classical PCA?

Here we briefly review certain problems and issues faced by classical PCA (Section 1.3 of Chapter 1). We see how the several methods we have surveyed in the literature attempt to address these problems.

1. *Addressing high computational complexity of PCA.* Neural Network based methods (Section 2.4) are suitable for high-dimensional data because they compute principal component directions incrementally *without computation of covariance matrix*. EM algorithms also learn eigenvectors adaptively without computation of covariance matrix (Section 2.6). Other methods which do not compute covariance matrix include Covariance-free Incremental PCA [176] and Simple PCA [120]. More recently, 2DPCA methods (Section 2.3) have become popular for image data. 2DPCA methods compute more compact covariance matrix which needs less computations than traditional computation. More interestingly, Feature partitioning based PCA (FP-PCA) methods (Section 2.2) consume less computational requirements by computing sub-covariance matrices using sub-patterns.

2. *Addressing poor performance with data of prominent local variations (local feature extraction).* Most of the existing PCA methods (except FP-PCA methods) are based on extracting features from whole patterns, which form the basis for global feature extraction. These methods may work well in some situations (where variations spread across entire pattern), however, may not perform well

when the variations are limited to a part or block of patterns. FP-PCA methods (Section 2.2) were proved to be superior over classical PCA methods when the variations are restricted to a part or block of a pattern (i.e when local variations are prominent). FP-PCA methods divide each pattern into sub-patterns or blocks and extract local principal components from these blocks. FP-PCA methods form the basis for local feature extraction.

3. *Addressing Small Sample Size (SSS) problem.* If the number of samples are small as compared to dimensionality of the patterns, the feature extraction itself may not be effective. Some of the methods which reduce the SSS problems include (i) FP-PCA methods (Section 2.2)– these methods divide each pattern into sub-patterns and feature extraction is done from these sub-patterns (blocks) instead of whole pattens. It is clear that sub-pattern dimensionality (u) is less than pattern dimensionality (d), therefore reducing SSS problem, (ii) 2DPCA methods (Section 2.3) reduce SSS problem by computing more compact $n \times n$ covariance matrix instead of huge $m.n \times m.n$ matrix (m, n are dimensions of an image matrix). Due to compact covariance matrix, the method requires to compute less number (i.e. n instead of traditional $m.n$) of principal component vectors.

4. *Addressing the problem of handling missing data and outliers.* Many robust methods such as Fuzzy logic based, Neural Network based methods, etc, are proposed to handle outliers data while computing principal components [143] [30] [68] [32] [61] [77] [18] [151]. For addressing the data with missing values many methods were proposed [62] [101] [18] [148].

5. *Addressing the issue of non-linear data.* PCA methods with *kernel trick* can effectively capture non-linearity (Section 2.5). Non Linear Neural Networks (NLNN) are also proved to be useful to capture non-linearity in the patterns (Section 2.4).

6. *Addressing the issue of choosing number of Principal Components.* Right choice of principal components influence the classifier performance as well as total amount of variance (structure) in the reduced data. Many methods are proposed (Section 2.8) in the literature to choose number of PCs.

2.12 What is the Problem We are Solving?

In our work, we investigate FP-PCA methods as reviewed in section 2.2. The motivation for our study on FP-PCA methods is due to their interesting and novel characteristics. In contrast to other PCA methods which are based on whole patterns, FP-PCA methods are unique in their approach, which are based on novel idea of *partitioning each pattern into sub-patterns (blocks) and extract local features.*

FP-PCA methods alleviate some of the crucial problems of classical PCA (i.e. whole-pattern based (global) PCA). FP-PCA methods have the following advantages over classical PCA: (i) Reduced computational complexity, (ii) Improved recognition/classification rate by local feature extraction if local variations are prominent, (iii) Reduced small sample size problem. However, FP-PCA methods as seen in the literature suffer from the following problems:

1. They do not retain the original essence (or merits) of classical PCA.

2. These methods purely perform local feature extraction (i.e. feature extraction is limited to a subset of original feature set (or limited to sub-patterns)). Therefore, FP-PCA methods may not perform well if there exists global variations (or strong correlations between local features extracted from the sub-patterns). Please note that classical PCA (global PCA or whole-pattern based PCA method) works well in this case.

3. Summarization of variance is not good because the entire covariance structure is not utilized which yields more number of local PCs resulting in less dimensionality reduction. Please note that classical PCA (global PCA or whole-pattern based PCA method) has good summarization of variance, because it makes use of entire covariance structure.

4. FP-PCA methods do not exploit two-dimensional structure of image data (Section 2.3) because they all use classical PCA in a region or block to extract local features. In classical PCA, each sub-image is formed as a feature vector, thus collapsing the two dimensional matrix structure of image.

From our analysis, we understand that FP-PCA methods and classical PCA methods (i.e. global PCA or whole-pattern based PCA methods) are complementary approaches. Put it in other way, FP-PCA methods work well in some cases, in which case classical PCA (global or whole-pattern base PCA) may not perform well and vice versa.

The *objectives* of our investigation presented in this thesis are given as follows.

2.12.1 Objectives of Our Investigation in this Thesis

1. From our study, we understand that there is no conceptual study of FP-PCA methods. Thus there is no conceptual basis to understand the nuances of these methods like (i) What are the issues that arise due to partitioning of the patterns, (ii) Is there any impact on dimensionality reduction because of partitioning of the patterns?.

   (a) Can we perform rigorous investigation on FP-PCA methods systematically and propose a generalized framework and issues?

   These concerns are addressed in *Chapter 3*.

2. To propose a novel FP-PCA method which alleviates the crucial problems of both (i) classical PCA and (ii) FP-PCA methods and exploits the strengths of both the methods.

   (a) Less computational complexity (as with FP-PCA methods)

   (b) Reducing SSS problem (as with FP-PCA methods)

   (c) Good classification performance (Good generalization ability) when local variations are dominant (as with FP-PCA methods)

   (d) Good classification performance (Good generalization ability) when global variations are dominant (as with classical PCA methods)

   (e) Less number of coefficients or components, that is good summarization of variation or high dimensionality reduction (as with classical PCA methods)

   These concerns are addressed in *Chapter 4*.

3. To propose some novel FP-PCA methods exclusively for image data (by taking ideas of using feature partitioning and matrix structure of image data), which show

   (a) Less computational complexity (as with FP-PCA methods and better than 2DPCA)

   (b) Reducing SSS problem (as with FP-PCA methods and better than 2DPCA method)

   (c) Good classification performance (Good generalization ability) when local variations are dominant (as with FP-PCA methods)

   (d) Good classification performance (Good generalization ability) when global variations are dominant

   These concerns are addressed in *Chapter 5*.

4. No theoretical study on FP-PCA methods is found in the literature. Can we perform a theoretical analysis of FP-PCA methods which brings out

   (a) Deeper insight of FP-PCA methods and establish links to classical PCA (global PCA)

   (b) General properties of FP-PCA methods

   These concerns are addressed in *Chapter 6*.

5. We are aware of that FP-PCA methods show their superiority as compared to classical PCA methods. Cluster analysis is a well established tool for data analysis in pattern recognition and data mining.

(a) Can we extend the ideas of feature partitioning to improve Cluster Analysis?

These concerns are addressed in *Chapter 7*.

6. We know that subspace classification is one of the traditional approaches in pattern recognition and integrates feature extraction and classification in a seamless fashion.

   (a) Can we extend the ideas of feature partitioning to improve subspace classification?

   These concerns are addressed in *Chapter 8*.

In this thesis, we do not address the following issues: (i) The issue of non-linear data, (ii) Choosing number of principal components and (iii) Handling missing values data and outliers.

2.13 Summary

In this section, we reviewed the state-of-the-art of PCA literature, which include FP-PCA methods, Two-dimensional PCA methods, Neural Network based PCA methods, KPCA methods, etc. Subsequently, we discussed how these PCA methods address the problems faced by classical PCA. In this work, we focus on study of FP-PCA methods. In subsequent Chapters (i) we propose a common framework for FP-PCA methods and identify issues to be addressed, (ii) we bring out some novel FP-PCA methods, which alleviate the problems faced by both the existing FP-PCA

methods and classical PCA (global PCA) methods, (iii) we establish general properties of FP-PCA methods by performing a theoretical analysis and (iv) we extend our feature partitioning ideas to cluster analysis and subspace classification.

In the next Chapter, we start our journey by proposing a framework which brings the existing FP-PCA methods under a common framework and identify the issues need to be addressed in this framework. This framework forms the basis for our work.