A Study on various Human Facial Feature Extraction Techniques in High Dimensional Spaces

Jaimin H. Jani, Dr. Subhaschandra Desai

Abstract-In today's era where one's face is used for ease of access for permitted levels of access in either physical or logical-way, it's a very challenging task for the devices equipped with various hardware and software tools to perform such kind of job with desirable accuracy in real time. Feature extraction is a very crucial and important task in facial recognition. In this paper various feature extraction techniques in high dimensional spaces are discussed. The objective of this study is to investigate pattern recognition methods for high-dimensional sample spaces.In a real time scenario and from a performance perspective, the dimensionality could be one of the culprits and makes a significant impact on the effectiveness of the outcome. If the data is transformed to a lower dimensional space by finding a new axis-system in which most of the data variance is preserved in a few dimensions. This reduction may also have a positive effect on the quality of similarity for certain data domains such as text. Our analysis also indicates currently accepted techniques and impact on overall performance as far as the feature extraction phase of facial recognition is concerned.

INTRODUCTION

Face recognition is an active research area with a wide range of applications in the real world. In recent years, a defined face recognition pipeline, consisting of four steps i.e. detection, alignment, representation, and classification has been presented.

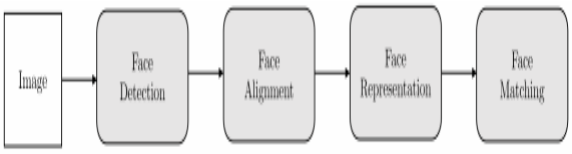

Fig. 1: Face recognition building blocks.

In the detection step the place of the image including face is found. The alignment step ensures the detected face is lined up with a target face or a model. In the representation step the detected face is described in a way that several descriptions with certain aspects about the detected face are presented. Finally, the classification step determines whether a certain feature corresponds with a target face or a model. Face recognition techniques are divided into Geometric and Photometric approaches. Geometric approaches consider individual features such as eyes, nose, mouth and a shape of the head and then develop a face model based on the size and the position of these characteristics. In photometric approaches the statistical values are extracted, subsequently, these values are compared with the related templates. A large number of researches have been devoted to feature extraction based on Gabor filter. A face representation using the Gabor filter, has been of focal importance in the machine vision, image processing and pattern recognition. In face recognition, the feature representation of a face is a critical aspect. If the representation step does not perform well, even the best classifiers cannot produce appropriate results. Good representations are those that on one hand minimize intra-person dissimilarities, on the other hand maximize differences between persons. Additionally, a significant representation should be fast and compact. There are several views related to the classification of the feature extraction methods. One possible classification divides the feature extraction methods into Holistic Methods and Local Feature-based Methods. In the first method the whole face image is applied as an input of the recognition operation similar to the well-known PCA-based method which was used in Kiby and Sirovichfollowed by Turk and Pentland. In the second method local features are extracted, for example the location and local statistics of the eyes, nose and mouth are used in the recognition task. EBGM methods are included in this category. Lades suggested a face recognition system based on DLA (Dynamic Link Architecture) platform, using extracting Gabor jets from each node over the rectangular grid to recognize faces. Wiskottexpanded DLA and introduced EBGM (Elastic Base Graph) method based on a wavelet to recognize the face. However, both LDA and EBGM have a high computational cost. Although the Gabor filters are computationally expensive due to a high dimension of the feature vector the results obtained from them are robust. T.Ojalaintroduced an original LBP operator which is regarded as a strong tool for describing the image texture.

Due to digitization, a huge volume of data is being generated across several sectors such as healthcare, production, sales, IoT devices, Web, organizations. Machine learning algorithms are used to uncover patterns among the attributes of this data. It has been demonstrated that high-dimensional space is significantly different from the three-dimensional (3-D) space, and that our experience in 3-D space tends to mislead our intuition of geometrical and statistical properties in high-dimensional sample spaces.

1.Characteristic Properties of High- Dimensional Spaces

For a fixed number of training samples, increasing the dimensionality of the sample space spreads the data over a greater volume. This process reduces overlap between the classes and enhances the potential for discrimination. Therefore, it is reasonable to expect that high dimensional sample spaces contain more information of capability to detect more classes with more accuracy. However, from the curse of dimensionality, we know that there is a penalty in classification accuracy as the number of features increases

beyond some point. Therefore, techniques of carrying out computations at full dimensionality may not deliver the advantages of high-dimensional sample spaces if there are insufficient training samples.

Experiments have shown that high-dimensional sample spaces are mostly empty since data typically concentrate in an outside shell of the sample space far from the origin as the dimensionality increases. This implies that the data samples are usually in a lower dimensional structure. As a consequence, high-dimensional data can be projected to a lower dimensional subspace without losing significant information in terms of separability among the classes by employing some feature extraction techniques. It has been also proved that as the dimensionality of the sample space goes to infinity, lower-dimensional linear projections approach a normality model with a probability approaching one. Here normality implies either a normal or a mixture of normal distributions. It turns out that the normally distributed high-dimensional data concentrate in the tails and uniformly distributed high-dimensional data concentrate in the corners. This makes density estimation task for high-dimensional sample spaces a difficult task. In this case, local neighborhoods become empty, which in turn produces the effect of losing detailed density estimation.

Another interesting observation was related to the first and the second order statistics of data samples. It has been shown that for low-dimensional sample spaces, class means representing first order statistics play a more important role in discriminating between classes than the class covariances representing second order statistics. However, as dimensionality increases, class covariance differences become more important.

In summary, the dimensionality of the sample space must be reduced before the application of the classifier to data samples in high-dimensional sample spaces. However, in order to keep the discriminatory information, which the high-dimensional sample spaces provide, good dimension reduction techniques are needed. In this study, the dimension reduction techniques for high-dimensional sample spaces are investigated.

2.DIMENSIONALITY REDUCTION

Dimensionality reduction usually improves the accuracy of recognition of a pattern recognition system besides saving memory and time consumptions. This seems somewhat paradoxical since dimensionality reduction usually reduces the information content of the input data. However, a good dimensionality reduction technique keeps the features with the high discriminative information and discards the features with redundant information. Thus, the worst effects of the curse of dimensionality are reduced after the dimensionality reduction process, and often improved performance is achieved over the application of the selected classifier in the original sample space. But given a set of features, how can the best set of features for classification be selected? Given a set of features, selection of the best set of features can be achieved in two different ways.

The first approach is to identify the features that contribute most to class separability. Therefore, our task is the selection of previously decided features out of our initial *d* features. This is called feature selection. The second approach is to compute a transformation which will map the original input space to a lower-dimensional space by keeping the most of the discriminative information. This transformation can be linear or nonlinear combinations of the samples in the training set. This approach is usually called the feature extraction. Both approaches require a criterion function, *J*, which is used to judge whether one subset of features is better than another. Exploring high-dimensional data is central to many application domains such as statistics, data science, machine learning, and information visualization. The main difficulty encountered in this task is the large size of such datasets, both in the number of observations (also called samples) and measurements recorded per observation (also called dimensions, features, variables, or attributes)

FEATURE SELECTION

In this approach we select the best set of features for classification out of original *d* features. We must first define a criterion function, *J*, to accomplish this task. The selected criterion is evaluated for all possible combinations of features systematically selected from *d* features. Then, we select the set of features for which the criterion is maximum as our final features. However, this task is not very straightforward because there are

$$\frac{d!}{(d-\tilde{d})!\,\tilde{d}!}$$

possible combinations for evaluation. As a consequence, this procedure may not be feasible even for moderate values of *d* and therefore, we will not consider the feature selection methods in this study since we are only interested in the data sets with high-dimensional spaces.

3. FEATURE EXTRACTION

In this approach we seek a transformation which will map the original input space to a lower dimensional space by keeping the features offering high classification power. The optimization is evaluated over all possible transformations of the data samples. Let denote the sought transformation for which, where ϖ is the family of allowable transformations and *x* refers to the training set samples. The new samples in the transformed space are computed by $y = W(x)$. The criterion function is typically a measure of distance or similarity between training set samples.

Linear Feature Extraction Methods

Feature extraction has been one of the most important issues of pattern recognition. Most of the feature extraction literature has centered on finding linear transformations, which map the original high-dimensional sample space into a lower-dimensional space that hopefully contains all discriminatory information. As explained previously, the principal motivation behind dimensionality reduction by feature extraction is that it may reduce the worst effects of the curse of dimensionality. Also linear feature extractions techniques are often used as pre-processors before more complex nonlinear classifiers. In the following sections we discuss these linear methods.

Generally, the face recognition process is divided into 3 regions such as Holistic method use the original image as an input for the face recognition system. The examples for holistic methods are PCA, LDA, and ICA and so on. In the Feature based method, the local feature points such as eye, nose, and mouth are first extracted, then it will be sent to the classifier. Finally, a Hybrid method is used to recognize both the local feature and whole face region. In Dimensionality reduction, Feature extraction is an important task to collect the set of features from an image. According to the author, Feature extraction or transformation is a process through which a new set of features is created. The feature transformation may be a linear or nonlinear combination of original features. This survey provides some of the important linear and nonlinear techniques listed as follows.

3.1 Principal Component Analysis (PCA)

PCA is one of the popular techniques for both dimensionality reduction and face recognition since the 1990's. Eigenfacesbuilt on the PCA technique is introduced by M.A.Turk and A.P.Pentland. It is a holistic approach where the input image is directly used for the process. PCA algorithm can be used to find a subspace whose basis vectors correspond to the maximum variance directions in the original *n* dimensional space. PCA subspace can be used for presentation of data with minimum error in reconstruction of original data. More survey papers provide the information for PCA techniques. MPCA and KPCA are fully based on the PCA technique.

3.2 Linear Discriminant Analysis (LDA)

LDA is one of the most famous linear techniques for dimensionality reduction and data classification. The main goal of the LDA consists in finding a base of vectors providing the best discrimination among the classes, trying to maximize the between-class differences, minimizing the within-class ones by using scatter matrices. It also suffers from a small sample size problem which exists in high dimensional pattern recognition tasks where number of available samples are smaller than dimensionality of the samples. DLDA, R-LDA, and KDDA are variations of LDA. This technique is also discussed in more survey papers.

3.3 Singular Value Decomposition (SVD)

SVD is an important factor in the field of signal processing and statistics. it is the best linear dimensionality reduction technique based on the covariance matrix. The main aim is to reduce the dimension of the data by finding a few orthogonal linear combinations of the original variables with the largest variance. Most of the researchers have also used this technique for face recognition.

3.4 Independent Component Analysis (ICA)

ICA is a statistical and computational technique for enlightening the hidden factors that underlie sets or random variables, measurements, or signals. ICA is superficially related to principal component analysis and factors analysis. The ICA algorithm aims at finding S components as independent as possible so that the set of observed signals can be expressed as a linear combination of statistically independent components. It use cosine measures to perform the covariance matrix and also it is better than the PCA and LDA performance.

3.5 locality Preserving Projections (LPP)

LPP can be seen as an alternative to Principal Component Analysis (PCA). When the high dimensional data lies on a low dimensional manifold embedded in the ambient space, the Locality Preserving Projections are obtained by finding the optimal linear approximations to theEigen functions of the Laplace Beltrami operator on the manifold. As a result, LPP shares many of the data representation properties of nonlinear techniques such as Laplacian Eigenmapsor Locally Linear Embedding.

3.6 multi Dimensional Scaling (MDS)

Multidimensional Scaling (MDS) is a linear Model for dimensionality reduction. MDS generates low dimensional codes placing emphasis on preserving the pairwise distances between the data points. If
the rows and the columns of the data matrix D both have mean zero, the projection produced by MDS will be the same as that produced by PCA. Thus, MDS is a linear Model for dimensionality reduction having the same limitations as PCA.

3.7 partial Least Squares

Partial least squares is a classical statistical learning method. It is widely used in chemo metrics and bioinformatics etc. In recent years, it is also applied in face recognition and human detection. It can avoid the small sample size problem in linear discriminant analysis (LDA). Therefore it is used as an alternative method of LDA.

4. NON LINEAR FEATURE EXTRACTION OF DIMENSIONALITY REDUCTION TECHNIQUES

This section presents a general introduction to nonlinear feature extraction methods employing kernel functions. The kernel trick concept has been introduced here, and this trick is applied to the linear DCV Method to make it a nonlinear method.

Non-linear methods can be broadly classified into two groups: a mapping (either from the high dimensional space to the low dimensional embedding or vice versa), it can be viewed as a preliminary feature extraction step and visualization is based on neighbor's data such as distance measurements. Research on non-linear dimensionality reduction methods has been explored extensively in the last few years.

4.1 An Introduction to Kernel Feature Extraction Methods

Sometimes linear methods may not provide sufficient nonlinear discriminant power for classification of linearly non-separable classes (e.g., exclusive-or problem). Thus, kernel methods have been proposed to overcome this limitation. The basic idea of these methods is first to transform the data samples into a higher-dimensional space ℑ via nonlinear mapping φ (.), and then apply the linear methods in this space. More formally, we apply the mapping

$$\varphi : Rd \rightarrow \Im,\ x\ a\varphi\ (x)$$

to all the data samples. The motivation behind this process is to transform linearly non-separable data samples into a higher-dimensional space where the data samples are linearly separable as illustrated in Figure 4.1. Since the mapped space is nonlinearly related to the original sample space, nonlinear decision boundaries between classes can be obtained for classification. This approach seems to contradict the curse of dimensionality phenomenon since it increases the dimensionality of the sample space for a fixed number of available training set samples. A satisfactory explanation for this dilemma lies in statistical learning theory. This theory tells us that learning in high-dimensional space can be simpler if one uses low complexity, i.e., a simple class of decision rules such as linear classifiers. In other words, it is not the dimensionality but the complexity of the function that matters. In some recognition tasks we may have sufficient knowledge about the problem and can choose φ (.) by hand. If the mapping is not too complex and is not too high-dimensional, we can explicitly apply this mapping as happens in Radial Basis Networks or Boosting Algorithms. However, in most cases we may not have sufficient prior knowledge to design φ (.), or the mapping of the data samples into a higher-dimensional space explicitly cannot be intractable. In such cases, we utilize kernel functions to circumvent these limitations.

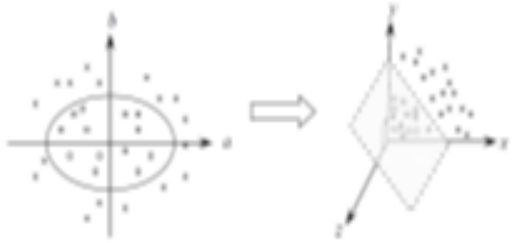

Figure 2: Kernel (nonlinear) mapping of 2-dimensional data into 3-dimensional space by polynomial kernel function.

In the following, a brief introduction to several non-linear dimensionality reduction techniques will be given.

4.1.1 Kernel Principal Component Analysis (KPCA)

Kernel PCA (KPCA) is the reformulation of traditional linear PCA in a high-dimensional space that is constructed using a kernel function. In recent years, the reformulation of linear techniques using the 'kernel trick' has led to the proposal of successful techniques such as kernel ridge regression and Support Vector Machines. Kernel PCA computes the principal eigenvectors of the kernel matrix, rather than those of the covariance matrix. The reformulation of traditional PCA in kernel space is straightforward, since a kernel matrix is similar to the in product of the data points in the high-dimensional space that is constructed using the kernel function. The application of PCA in kernel space provides Kernel PCA the property of constructing nonlinear mappings.

4.1.2 Isometric Mapping (ISOMAP)

Most of the linear methods do not take the neighboring data points into an account. ISOMAP is a technique that resolves this problem by attempting to preserve pair wise geodesic (or curvilinear) distances between data points. The approximation of geodesic distance is divided into two cases. For neighboring points, Euclidean distance in the input space provides a good approximation to geodesic distance and faraway points, geodesic distance can be approximated by adding up a sequence of "short hops" between neighboring points. ISOMAP shares some advantages with PCA, LDA, and MDS, such as computational efficiency and asymptotic convergence guarantees, but with more flexibility to learn a broad class of nonlinear manifolds.

4.1.3 Locally Linear Embedding

Locally linear embedding (LLE) is another approach which addresses the problem of nonlinear dimensionality reduction by computing low dimensional, neighborhood preserving embedding of high-dimensional data. It is a technique that is similar to ISOMAP in that it also constructs a graph representation of the data points. It describes the local properties of the manifold around a data point x_i by writing the data point as a linear combination w_i (the so-called reconstruction weights) of its k nearest neighbors x_{ij} and attempts to retain the reconstruction weights in the linear combinations as good as possible.

4.1.4 Laplacian Eigenmaps:

A closely related approach to locally linear embedding is Laplacian eigenmaps. Given *t* points in *n*-dimensional space, the Laplacian eigenmaps Method (LEM) starts by constructing a weighted graph with *t* nodes and a set of edges connecting neighboring points. Similar to LLE, the neighborhood graph can be constructed by finding the *k* nearest neighbors. The final objectives for both LEM and LLE have the same form and differ only in how the matrix *is* constructed.

4.1.5 Stochastic Neighbor Embedding:

Stochastic Neighbor Embedding (SNE) is a probabilistic approach that maps high dimensional data points into a low dimensional subspace in a way that preserves the relative distances to near neighbors. In SNE, similar objects in the high dimensional space will be put nearby in the low dimensional space, and dissimilar objects in the high dimensional space will usually be put far apart in the low dimensional space. A Gaussian distribution centered on a point in the high dimensional space is used to define the probability distribution that the data point chooses other data points as its neighbors. SNE is superior to LLE in keeping the relative distances between every two data points.

4.1.6 Semi Definite Embedding (SDE):

Semi definite Embedding (SDE), can be seen as a variation of KPCA and an algorithm is based on semi definite programming. SDE learns a kernel matrix by maximizing the variance in feature space while preserving the distances and angles between nearest neighbors. It has several interesting properties: the main optimization is convex and guaranteed to preserve certain aspects of the local geometry; the method always yields a semi positive definite kernel matrix; the eigenspectrum of the kernel matrix provides an

estimate of the underlying manifold's dimensionality; also, the method does not rely on estimating geodesic distances between far away points on the manifold. This particular combination of advantages appears unique to SDE.

5. CONCLUSION

Because of varying applications and span over different domains, selection of appropriate feature extraction techniques make a major impact in computation required (i.e. time and space complexity) in face recognition. Scholars have conducted and explored various aspects vigorously in this area for the past many years, and though significant amounts of progress has been achieved so far. Feature extraction is one of the most preprocessing and fundamental task in face recognition tasks. This paper contained a detailed survey on various existing feature extraction techniques for face recognition. Different face recognition algorithms can be applied on available databases. Even when the same database is used, researchers may use different protocols for testing. After a detailed review of a number of research papers, we found two main points (1) For the best-performing supervised defect prediction models, correlation and consistency-based feature selection techniques should be appropriate and (2) Neural network-based feature reduction techniques generate features that have a small variance across both supervised and unsupervised defect prediction models. In summary, a face recognition system should not only be able to cope with variations in illumination, expression and pose, but also recognize a face in real-time. We recommend that practitioners who do not wish to choose a best-performing defect prediction model for their data use a neural network-based feature reduction technique.

REFERENCES

[1] Ngoc-Son Vu, H. M. Dee and A. Caplier, (2012) "Face recognition using the POEM descriptor", Pattern Recognition.
[2] C. Liu and H. Welchsler, (2001) "Gabor feature classifier for face recognition", in processing of the ICCV, Vol. 2, No. 5, pp 270-275.
[3] J.R. Movellan, "Tutorial on Gabor filters",http://mplab.ucsd.edu/tutorials/gabor.pdf.
[4] M. Zhou, and H. Wei, (2006) "Face verification using Gabor Wavelets and AdaBoost", 18th International Conference on Pattern Recognition, pp 404-407.
[5] M.Kirby and L. Sirovish, (1990) "Application of the Karhunen-Lo_ve procedure for the characterization of human faces", IEEE Transactions on Pattern Analysis and Machine Intelligence12, pp 103-108.
[6] M.Turk and A.P. Pentland, (1991) "Eigen faces for recognition", Journal of Cognitive Neuroscience, pp 71-86.
[7] C. Aguerrebere, G. Capdehourat, M. Delbracio, M. Mateu, A. Fern´andez and F. Lecumberry, (2007) "Aguar´a: An Improved Face Recognition Algorithm through Gabor Filter Adaptation", Automatic Identification Advanced Technologies.
[8] M.Lades, J.C.Vorbruggen, J.Buhmann, J.Lang, C.V.Malsburg, C.Wurtz and W.Konen, (1993) "Distortion invariant object recognition in tha dynamic link architecture", IEEE Trans.Computers, Vol.42, No.3, pp 300-311.
[9] L.Wiskott, J.M.Fellous, N.Kruger, and C.VMalsburg, (1997) "Face recognition by elastic bunch graph matching, IEEE Trans, Pattern Aal. Match.Intel., Vol.19, No.7, pp 775-779.
[10] A. Bayesian, and C.H. Liu,(2007) "On Face Recognition using Gabor Filters", World Academy of Science Engineering and Technology 28, pp 51-56.
[11] T. Ojala, Pietikinen and Menp, (2002) "Multi resolution gray-scale and rotation invariant texture classification with local binary patterns", IEEE Transaction on Pattern Analysis and Machine Intelligence, pp 971-987.
[12] Jimenez, L. O. and Landgrebe, D. A. (1998) Supervised classification in high dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 28(1), 39-54.
[13] Veerabhadrappa, LalithaRangarajan," Bi-level dimensionality reduction methods using feature selection and feature extraction" International Journal of Computer Applications (0975 – 8887) Volume 4 – No.2, July 2010.
[14] Rama Chellappa, Charles L. Wilson, And SaadSirohey,"Human and machine recognition of faces: A Survey",Proceedings of the IEEE,1995
[15] William A. Barrett," A Survey of Face Recognition Algorithms and Testing Results",Proceedings of the IEEE,1998.
[16] W.Zhao,R.Chellapa,A.Rosenfield,P.J.Philips," Face Recognition : A Literature Survey",2001
[17] W.Zhao,R.Chellapa,A.Rosenfield,P. J.Philips," Face Recognition : A Literature Survey",ACM proceedings,2003
[18] XiaoyangTana,b, SongcanChena,c,∗, Zhi-Hua Zhoub, FuyanZhangb," Face recognition from a single image per person:Asurvey",Published in Elseiver,2006
[19] Patil A.M., Kolhe S.R. and Patil P.M," 2D Face Recognition Techniques: A Survey",2010
[20] M. Turk and A. Pentland, "Eigenfaces for recognition", Journal of Cognitive Neuroscience, vol. 3, No. 1, 1991, pp.71 - 86.
[21] S.K.Sandhu, SumitBudhiraja," Combination of Nonlinear Dimensionality Reduction Techniques for Face Recognition System",published in IJERA
[22] S.Sakthivel," enhancing face recognition using improved dimensionality reduction and feature extraction algorithms –an evaluation with orl database" international journal of engineering science and technology,2010
[23] Shylaja S S, K N Balasubramanya Murthy and S Natarajan," Dimensionality Reduction Techniques for Face Recognition", (IJACSA) International Journal of Advanced Computer Science and Applications, 2011
[24] Yunfei Jiang and Ping Guo," Comparative Studies of Feature Extraction Methods with Application to Face Recognition"IEEE,2007
[25] Ion Marqu´es," Face Recognition Algorithms",2010.
[26] CHEN Cai-ming, Zhang Shi-qing,ChenYuefen," Face Recognition Based on MPCA", 2nd International Conference on Industrial Mechatronics and Automation,2010
[27] Weilin Huang and Hujun Yin," linear and nonlinear dimensionality reduction for face recognition",IEEE,2009
[28] Schölkopf, B. and Smola, A. J. (2002) Learning with Kernels. MIT Press.
[29] Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K. and Schölkopf, B. (2001) An introduction to kernel-based learning algorithms. IEEE Transaction on Neural Networks, 12, 181-201.
[30] Ali Ghodsi," Dimensionality Reduction A Short Tutorial",2006
[31] Renqiang Min," A Non-linear Dimensionality Reduction Method for Improving Nearest Neighbour Classification",2005
[32] Thippa Reddy Gadekallu,Praveen Kumar Reddy,KuruvaLakshman,RajeshKaluri, "Analysis of Dimensionality Reduction Techniques on Big Data",2020
[33] MateusEspadoto, Rafael M. Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea, "Towards a Quantitative Survey of Dimension Reduction Techniques",2019