VOL. 13, NO. 23, DECEMBER 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

INTELLIGENT TECHNIQUES (LINEAR AND NONLINEAR) FOR VECTORS SIZE REDUCTION IN FEATURE SELECTION POINT

Mohamed A. El-Sayed[1,2]
[1]Department of Computer Science, College of Computers and IT, Taif University, Hawia, Taif, Kingdom of Saudi Arabia
[2]Department of Mathematics, Faculty of Science, Fayoum University, Fayoum, Egypt
E-Mail: drmsayed@yahoo.com

ABSTRACT

The paper will present the most known techniques for feature selection and size reduction of vectors. Some linear and nonlinear techniques are designed and implemented such as Kernel Principal Component Analysis, Locally Linear Embedding, MPPCA, Generalized Discriminant Analysis, Laplacian Eigen-maps, Isomap, Landmark Isomap, and LTSA approaches. These approaches are applied and tested on different common biometric database, such as DRIONS, VARIA, IIT Delhi Ear and STORE datasets. The experimental results of the suggested techniques are presented and compared.

Keywords: feature selection, biometric recognition, feature reduction techniques, dimensionality reduction, vectors size reduction.

1. INTRODUCTION

Biometric recognition indicates the automated recognition of users depends on their physiological and behavioral characteristics such as face, retina, iris, ears, voice, gait, fingerprint and vein [1, 2]. Biometrics are automated techniques of identifying a person or verifying the identity of a person based on a physiological or behavioral characteristic. Dynamic signature verification, speaker verification and keystroke dynamics are examples of behavioral characteristics.

The motivation behind the work: A wide variety of systems require reliable personal recognition schemes to either confirm or determine the identity of an individual requesting their services. The purpose of schemes is to ensure that the rendered services are accessed only by a legitimate user, and not anyone else. The hike in credit card fraud and identity theft in recent years indicates that authentication is an issue of major concern in wider society. Individual passwords, pin identification or even token based arrangement all have deficiencies that restrict their applicability in a widely-networked society.

Every biometric system is composed of four main phases, sensor phase, dimensionality reduction phase, matching phase and decision phase. See Figure-1.

In sensor phase, a sensor acquires the biometric characteristic of an individual and makes a digital description of it. In feature extraction phase, input sample is processed and generates a compressed vector called template. Template is stored in database or in a smart card. Feature selection / reduction techniques may be employed at this level to reduce the size of the extracted feature vectors. In the third phase, it compares the presented biometric sample with the stored templates. In final phase, depending on the matching score or security threshold, the system accepts or rejects the user. Generally, there are three main steps are introduced to dimensionality reduction: features selection step, classification step and optimization step.

In the system have been studying several methods to resolve the complexity process of the identification mode related to the dimensionality problem. The problem of slowness resulting from increased features used and the number of individuals. These methods achieve several objectives: first objective, reduce features that are used by reducing the space of vectors as possible. Second objective, clusters number of feature identification are raised to increase the system performance.

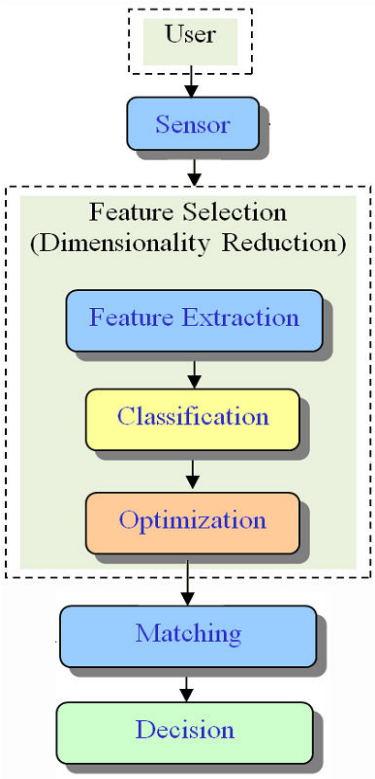


Figure-1. Four main phases for biometric system.

Nandakumar *et al* [3] used score level fusion multi-biometrics system by combining three traits (face, fingerprint and hand geometry) are presented, using compare for the feature extraction in each single trait.

Signature verified using the dynamic programming technique of string matching. Voice is verified using a commercial, off the shelf, software development kit. In order to improve the authentication performance, the paper combines information from both on-line signature and voice biometrics. fusion is performed at the matching score level in S. Krawczyk [4]. Connaughton *et al* [5] present the multi-biometrics system exploits the face information, a sensor that is intended for iris recognition purposes, with no modifications to the sensor and no increase in probe data acquisition time. The resulting system is less likely to experience failures to acquire, and the use of multiple modalities could allow the system to identify subjects with incomplete gallery data. This approach could be extended to operate on other stand-off iris sensors, which often detect the face as a preliminary step to iris image acquisition. Kumar *et al*. [6] applied palm print and hand geometry over other biometric modalities. It implemented particle swarm based optimization technique for selecting optimal parameters through decision level fusion of two modalities: palm print and hand geometry. Vanaja *et al* [7] applied another paper in biometrics, it used multi-biometrics for security. Security systems having realized the value of biometrics for two basic purposes: to identify users. It focuses on an efficient methodology for identification and verification for iris detection.

Muthukumar *et al*. [8] focused on the security of the biometric system, and proposed a multimodal system based on an evolutionary algorithm, Particle Swarm Optimization that adapts for varying security environments. With these two concerns, this paper had developed a design incorporating adaptability, authenticity and security.

Soviany *et al* [9] proposed a classification method for people identification accuracy improvement in which the biometric system is trained not for all enrolled individuals and reducing the computational complexity for the large-scale biometric identification. The biometric detectors are relying on non-linear models which are more suitable for the real biometric data with high degree of intra-class variance; therefore they improve the people recognition accuracy even for the most difficult cases.

Feature selection in S. F. Pratama *et al* [10] is used to obtain the unique individual significant features which are proven very important in handwriting analysis of Writer Identification domain.

El-Sayed *et al*. [11] proposed identity verification technique of individuals based on retinal features, It an optimized feature template using teacher learning based optimization process. The process of feature extraction performed by Gabor wavelet transform function, these wavelet transform function basically texture feature extraction technique.

Tomassi *et al*. [12] proposed several approaches of varying computational complexity for analyzing multiple correlated markers that are also censored due to lower and/or upper limits of detection, using likelihood-based sufficient dimension reduction (SDR) methods. They find that explicitly accounting for the censoring in the likelihood of the SDR methods can lead to appreciable gains in efficiency and prediction accuracy, and also outperformed multiple imputations combined with standard SDR.

Liu *et al.* [13] proposed a feature selection method based LW-index statistical measure; the measure replaced the expensive cross-validation scheme to evaluate the feature subset. This index combined with Sequence Forward Search algorithm (SFS-LW) and proposed method obtained similar classification accuracy as the wrapper method with centroid-based classifier or SVM.

Istitieh and El-Sayed [14] studied method depend on biometric features, entropy properties and dimensionality reduction. The method consist of two main phases, first phase, we will construct a biometric identification technique with multi-model features extraction. In second phase, the various properties of entropy applied on multi-model features extraction to increase the accuracy factor of security recognition in this system.

El-Sayed and Nada [15] discussed the behavior of known projection indices optimized by genetic algorithm and particle swarm optimization algorithm to confirm the performance of the famous family SVM approaches. The experimental results of the indices techniques when applied on standard datasets using GA and PSO algorithms are provided then compared the results against the common methods to light the complexity and quality.

The reminder of this manuscript is organized as follows: Section 2 briefly explains the most known techniques for feature selection and size reduction of vectors. Experiments are designed and implemented of the most known techniques. Section 3 presents the simulation work and analysis of experiments after tested on different biometric database. Finally, conclusions are presented in Section 4.

2. FEATURE SELECTION AND VECTORS SIZE REDUCTION

Different dimensionality reduction techniques were implemented on feature dataset for dimensional reduction and the strength and weakness of techniques were compared in various articles. From literature survey it has been noted that many linear and nonlinear techniques were found to be more robust for expression recognition. The most known techniques are:

2.1 Principal Components Analysis (PCA)

PCA [16-19] is a very common mechanism for dimensionality reduction. PCA reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables. PCA finds a low-dimensional linear subspace such that when x is projected there information loss (here denoted as variance) is minimized. Finds directions of maximal variance. Equivalent to finding eigenvalues and eigenvectors of the covariance matrix. Suppose that the set d principal axes of data vectors $x_1, x_2, ..., x_t$ and t is the observations in an

$n \times t$ matrix X, each column corresponds to an n-dimensional observation. The main steps of the algorithm:

a) Basis of recover: compute $XX^T = \sum_{i=1}^{t} x_i x_i^T$ and suppose that U is eigenvectors of XX^T corresponding to the top d eigenvalues.

b) Set training data in encoded form: let $d \times t$ matrix encodings Y of the original data, $Y = U^T X$.

c) Training data is remodel according to $\hat{X} = UY = UU^T X$.

d) Test example in encoded form: y is a d-dimensional encoding of x, since $y = U^T x$.

e) Test example is reconstruct according to $\hat{x} = Uy = UU^T x$.

2.2 Locally Linear Embedding (LLE)

LLE [20-22] is another mechanism to solve the problem of nonlinear dimensionality reduction. The high-dimensional space converts to the system of single global coordinate such that the relationships among neighboring points in space are kept. This approach concentrate on three steps:

a) For every data point in X_i, the neighbors are identified. This can be accomplished by getting the k nearest neighbors of X_i, or by selecting all points inside some specified radius ε.

b) Calculate the weights W_{ij} that gives best linearly reconstruct each x_i from its neighbors. By minimizing the cost function $\min_W \| X_i - \sum_{i=1}^{K} W_{ij} X_j \|^2$ by constrained linear fits W_{ij}.

c) Using the weights W_{ij} determined in the step 2, we get the best reconstructed of low-dimensional embedding vectors Y_i by minimizing $\min_Y \sum_{i=1}^{N} \| Y_i - YW_i \|^2$.

2.3 Kernel Principal Components Analysis (KPCA)

Traditional PCA applies linear transformation; it may not be effective for nonlinear data. KPCA is a solution of this problem, apply map of nonlinear transformation to potentially very high-dimensional data points, For computational efficiency, we can apply the kernel trick by rewritten PCA in terms of dot product, For more details you can see [23-25].

2.4 Nonlinear techniques include mapping subspace (Isomaps)

Isomaps produces globally optimal low-dimensional Euclidean representation of nonlinear highly curved input space. This approach concentrates on three key steps:

a) In high-dimensional data space, rebuild neighborhood graph G such as LLE.

b) For each pair of points in G, Computing geodesic pairwise distances through shortest path distances.

c) Use classical multidimensional scaling; introduced by Borg and Groenen, 1997; as to preserve these distances with geodesic distances. Then convert Euclidean distance to Geodesic distance between all pairs of points in G. For more details you can see [26, 27].

2.5 Laplacian Eigenmaps (LEigen)

This approach [28-30] concentrates on four key steps:

a) Construct the adjacency graph G, such that if x_i and x_j are closed and satisfy $\| x_i - x_j \|^2 < \varepsilon$ then create an edge between verities i and j. Also the adjacency graph can rebuild by nearest neighbors approach.

b) Choose weights for edges in the graph using the heat kernel similarity, if x_i and x_j are connected then $W_{ij} = \exp[-\| x_i - x_j \|^2 / t]$, $t \in \Re$. if parameter t is not determine, put $W_{ij} = 1$ of connected pair (x_i, x_j).

c) Let G is connected, calculate eigenvalues and eigenvectors of the equation $Lf = \lambda Df$. Where D is diagonal of symmetric weight matrix $D_{ii} = \sum_{j=1} W_{ji}$, and Laplacian matrix $L = D - W$.

d) The low-dimensional embedding, suppose $f_0, f_1, ..., f_{k-1}$ be the solution and their eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq ... \leq \lambda_{k-1}$. After negligible of f_0 corresponding to $\lambda_0 = 0$, the reset eigenvectors embedded in m-dim Euclidean space $x_i \rightarrow (f(i)_1, ..., f_m(i))$.

3. EXPERIMENTAL AND RESULTS

3.1 Data sets

For experimentation, we used three various data sets related to user biometric identification and reduction dimension, like database about ear and retinal.

First database: DRIONS-DB [31, 32], Digital Retinal Images for Optic Nerve Segmentation database. The database consists of 110 colour digital retinal images; they were digitized using a HP-PhotoSmart-S20 high-resolution scanner, RGB format, resolution 600×400 and 8 bits/pixel. Initially, it were obtained 124 eye fundus images selected randomly from an eye fundus image base belonging to the Ophthalmology Service at Miguel Servet Hospital, Saragossa (Spain). From this initial image base, all those eye images (14 in total) that had some type of cataract (severe and moderate) were eliminated and, finally, was obtained the image base with 110 images. Figure-2 presents gray examples of eye fundus image in DRIONS Database.

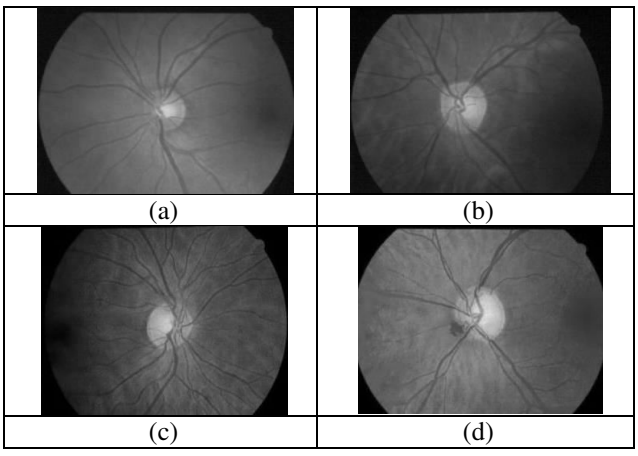


Figure-2. DRIONS database samples.

Second database: VARIA [33, 34], The database is a set of retinal images used for authentication purposes. The database currently includes 233 images, from 139 different individuals. The images have been acquired with a TopCon non-mydriatic camera NW-100 model and are optic disc centered with a resolution of 768×584. See Figure-3.

Third database: IIT Delhi Ear [35, 36], The database consists of the ear image database collected from the students and staff at IIT Delhi, New Delhi, India, see Figure-4. It is acquired from the 121 different subjects and each subject has at least 3 ear images. The database of 471 images has been sequentially numbered for every user with an integer identification/number. The resolution of these images is 272 × 204 pixels and cropped ear images of size 50 × 180 pixels.

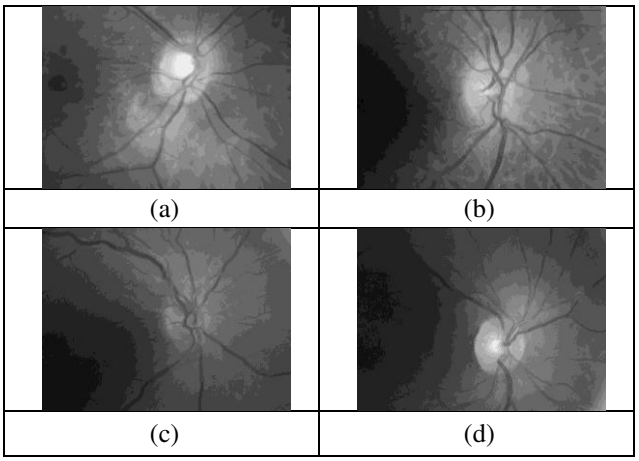


Figure-3. VARIA database samples.

Fourth database is the structured analysis of the retina (STARE) originally collected by Hoover *et al.* [37]. It consists of a total of twenty eye fundus color images where ten of them contain pathology. The images were captured using a TopCon TRV-50 fundus camera with FOV equal to thirty-five degree. Each image resolution is 700 × 605 pixels with eight bits per color channel and it's available in PPM format. The set of twenty images are not divided into separated training and testing sets. See Figure-5.

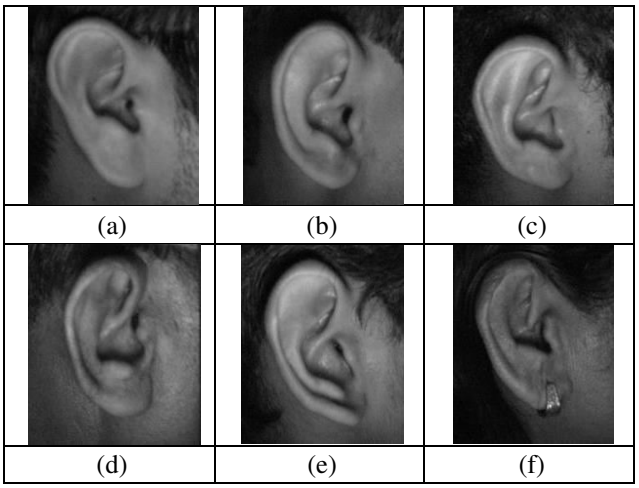


Figure-4. IIT Delhi ear database samples.

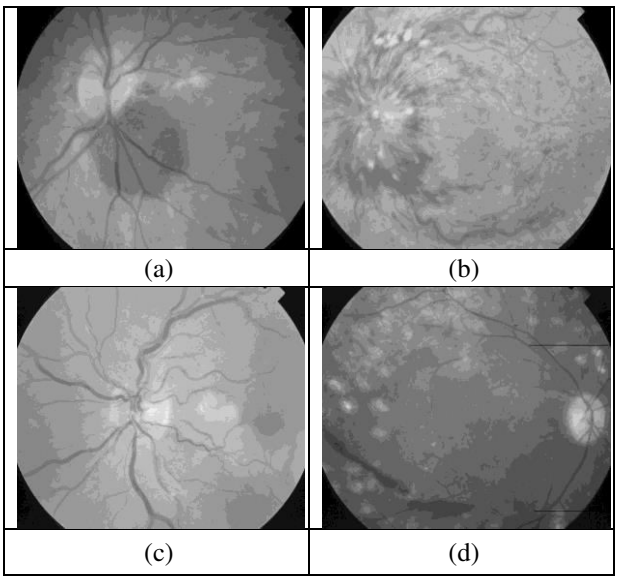


Figure-5. STARE database samples.

The various sizes and numbers of full/cropped images are used and summarized in Table-1.

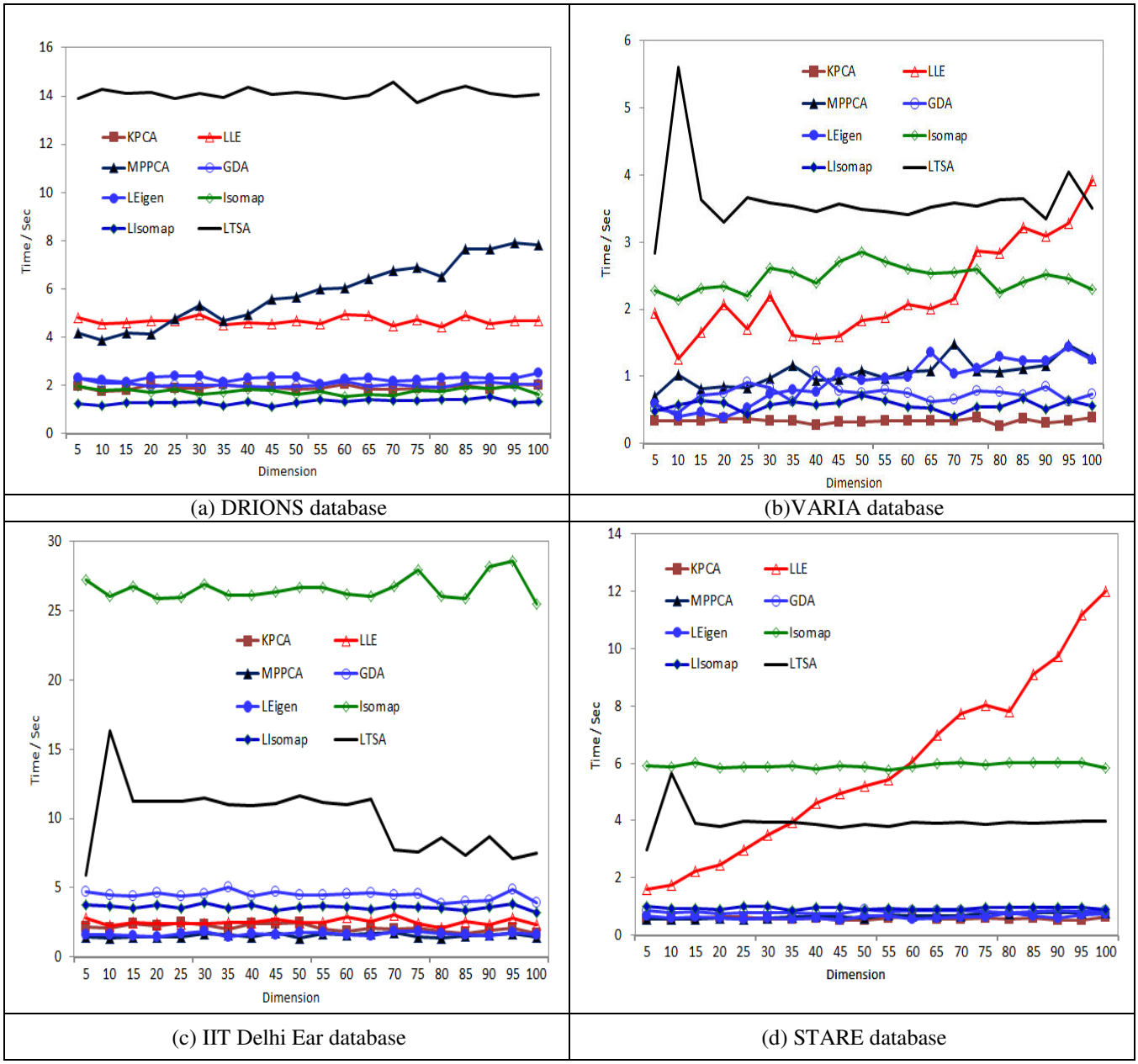
Table-1. Selected datasets for study of feature selection using common techniques.

Database	Images number	Image size	Vector length
DRIONS	110	600×400	240000
VARIA cropped	233	92×112	10304
IIT Ear cropped	793	50×180	9000
STARE	386	600×500	300000

3.2 Average time with various methods

We study 8 common dimensionality reduction algorithms (linear and nonlinear techniques), KPCA, LLE, MPPCA, GDA, LEigen, Isomap, Landmark Isomap, and LTSA approach. These algorithms are applied on four datasets, DRIONS, VARIA, IIT Delhi Ear and STORE databases and executed on a Dell -PC running Windows 7 professional with 2.4 GHz Intel Core i5 processor and 4GB RAM.

Figure-6 shows the required time of each method for dimension reduction of vector length of image. The approaches are tested when generate 5, 10,…or 100 features for each data sets in Table-1.



(a) DRIONS database (b)VARIA database

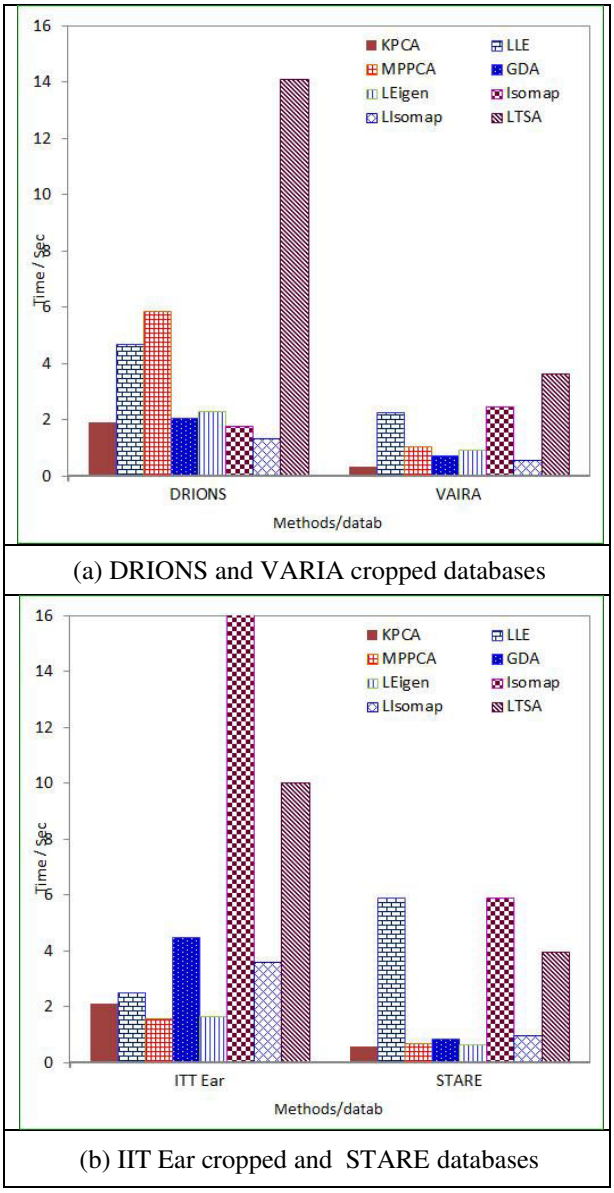
(c) IIT Delhi Ear database (d) STARE database

Figure-6. Various databases and run time for each method.

Here we present the various techniques to reduce the dimensions of the original datasets. From the survey, it comes to know that LDA, GDA, LLE and MPPCA are the powerful techniques to handle the linear types of data. KPCA, Landmark Isomap and LTSA are effectively worked on non-linear data while computing low and high dimensional feature dataset. The drawback of these nonlinear embedding techniques are consumes large time while computing high dimensional feature dataset such as Isomap and TSNE, see Figure-7. But the nonlinear techniques are efficient compared with the linear techniques of extraction of good features in the non-linear real world data.

4. CONCLUSIONS

This paper presents the most known techniques for feature selection and size reduction of vectors. Some linear and nonlinear techniques are designed and implemented such as KPCA, LLE, MPPCA, GDA, LEigen, Isomap, LIsomap, and LTSA approach. These approaches are applied and tested on different common biometric database, such as DRIONS, VARIA, IIT Delhi Ear and STORE datasets. The drawback of these nonlinear embedding techniques are consumes large time while computing high dimensional feature dataset such as Isomap and TSNE. But the nonlinear techniques are efficient compared with the linear techniques of extraction of good features in the non-linear real world data.



(a) DRIONS and VARIA cropped databases

(b) IIT Ear cropped and STARE databases

Figure-7. Average time of the techniques / database.

ACKNOWLEDGEMENT

The author is very pleased to offer the sincere thanks and appreciation to King Abdulaziz City for Science and Technology (KACST), which provided the grant No. 35-276. Also, the author thanks Taif University for its continuous support in the preparation of the project.

REFERENCES

[1] M. A. El-Sayed, S. F. Bahgat, and S. Abdel-Khalek. 2013. New approach for identity verification system using the vital features based on Entropy. International Journal of Computer Science Issues (IJCSI). 10(6): 11-17.

[2] Mohamed A. El-Sayed. 2015. Proposed System of Biometric Authentication Using Palm Print/Veins with Tsallis Entropy. International Journal of Computer Science and Technology. 6(2): 9-14.

[3] K. Nandakumar, Anil K. Jain and R. Arun. 2009. Fusion in Multibiometrics Identification Systems: What about the Missing Data? Proc. of ICB, Alghero.

[4] S. Krawczyk. 2005. User Authentication Using On-Line Signature and Speech. Mc.S thesis, Department of Computer Science and Engineering, Michigan State University.

[5] K.W. R. Connaughton and P. Flynn. 2007. Fusion of Face and Iris Biometrics‖ University of Notre Dame, IN 46637.

[6] A. Kumar, M. Hanmandlu, H. Sanghvi and H.M. Gupta. 2010. Decision level biometric fusion using Ant Colony Optimization. Image Processing (ICIP), IEEE International Conference on, on page(s): 3105 - 3108.

[7] R.E. Vanaja and E.R.Ch. Waghmare. 2011. Iris Biometric Recognition for Person Identification in Security Systems. International Journal of Computer Applications (0975 - 8887) Vol. 24, No.9.

[8] A. Muthukumar, C. Kasthuri and S. Kannan. 2012. Multimodal Biometric Authentication Using Particle Swarm Optimization Algorithm with Fingerprint and Iris. ICTACT Journal on Image and Video Processing. 02(03): 369-374.

[9] S. Soviany, C. Soviany, S. Puşcoci, M. Jurian. 2013. Detection Optimization for Biometric Applications with Non-Linear Classification Models. Journal of Bioinformatics and Biological Engineering. 1(1): 1-9.

[10] S. F. Pratama, A. K. Muda, Y. Choo and N. A. Muda. 2014. A New Swarm-Based Framework for Handwritten Authorship Identification in Forensic Document Analysis. Computational Intelligence in Digital Forensics: Forensic Investigation and Applications, Studies in Computational Intelligence, Springer International Publishing Switzerland. pp. 385-411.

[11] M. A. El-Sayed, M. Hassaballah, M. A. Abdel-Latif. 2016. Identity Verification of Individuals Based on Retinal Features Using Gabor Filters and SVM. Journal of Signal and Information Processing. 7: 49-59.

[12] Diego Tomassi, Liliana Forzani, Efstathia Bura, Ruth Pfeiffer. 2016. Sufficient dimension reduction for censored predictors, Biometrics, Version of Record online http://onlinelibrary.wiley .com/ doi/10.1111/biom.12556/ full#references: 9 Aug.

[13] Chuan Liu, Wenyong Wang, Qiang Zhao, Xiaoming Shen, Martin Konan. 2017. A new feature selection method based on a validity index of feature subset, Pattern Recognition Letters. 92: 1-8.

[14] Yarob A. M. Istitieh, Mohamed A. El-Sayed. 2018. Dimensionality Reduction of Biometric Features based on Entropy Properties. International Journal of Computer Science & Applications. 13(6): 3619-3623.

[15] Mohamed A. El-Sayed, Yasser Nada. 2018. On Projection Indices of Classification in Biometric Identification Problems. International Journal of Applied Engineering Research (IJAER). (13) 7, 4663-4666.

[16] Nawaf Hazim Barnouti. 2016. Face Recognition using PCA-BPNN with DCT Implemented on Face94 and Grimace Databases. International Journal of Computer Applications. 142: 8-13.

[17] Sasan Karamizadeh, Shahidan M. Abdullah, Azizah A. Manaf, Mazdak Zamani, Alireza Hooman. 2013. An Overview of Principal Component Analysis. Journal of Signal and Information Processing. 4, 173-175.

[18] Nandakishore Kambhatla Todd K. Leen. 1997. Dimension Reduction by Local Principal Component Analysis. Neural Computation. 9: 1493-1516.

[19] Archana H. Telgaonkar, Deshmukh Sachin. 2015. Dimensionality Reduction and Classification through PCA and LDA. International Journal of Computer Applications. 122(17): 4-8.

[20] Roweis S., Saul L. 2001. Nonlinear dimensionality reduction by locally linear embedding. In: IEEE ICCV. 290: 2323-2326.

[21] Deguang Kong, Chris Ding, Heng Huang, Feiping Nie. 2012. An Iterative Locally Linear Embedding Algorithm. Appearing in Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK.

[22] J. L. Ward1 and S. L. Lumsden. 2016. Locally linear embedding: dimension reduction of massive protostellar spectra, MNRAS. 461(2): 2250-2256.

[23] Van-Sang Ha, Ha-Nam Nguyen. 2016. C-KPCA: Custom Kernel PCA for Cancer Classification, Machine Learning and Data Mining in Pattern Recognition. Vol. 9729 of the series Lecture Notes in Computer Science. pp. 459-467.

[24] K. I. Kim, K. Jung, and H. J. Kim. 2002. Face recognition using kernel principal component analysis. IEEE Signal Processing Letters. 9: 40-42.

[25] A. Vinay, Vinay.S. Shekhar, K.N. Balasubramanya Murthy, S. Natarajan. 2015. Face Recognition Using Gabor Wavelet Features with PCA and KPCA - A Comparative Study. 3rd International Conference on Recent Trends in Computing 2015, Procedia Computer Science. 57: 650-659.

[26] Ashutosh Saxena, Abhinav Gupta, and Amitabha Mukerjee. 2004. Non-linear Dimensionality Reduction by Locally Linear Isomaps , Springer-Verlag Berlin Heidelberg, ICONIP 2004, LNCS 3316, pp. 1038-1043.

[27] Dipa Patra, Jnyanaranjan Dash, Chittaranjan Pradhan, R. N. Ramakant Parida. 2015. Mathematical Analysis of PCA, MDS and ISOMAP Techniques in Dimension Reduction, International Journal of Advance Research in Computer Science and Management Studies. 3(5): 61-67.

[28] M. Belkin, P. Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Advances in Neural Information Processing System, 15.

[29] M. Belkin, P. Niyogi. 2003. Laplacian eigenmaps for dimensionality. Speech Communication, 1(2-3): 349-367.

[30] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet. 2003. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. Mij, 1, 2.

[31] E.J. Carmona, M. Rincón, J. García-Feijoo and J. M. Martínez-de-la-Casa. 2008. Identification of the optic nerve head with genetic algorithms. Artificial Intelligence in Medicine. 43(3): 243-259.

[32] DRIONS-DB: Digital Retinal Images for Optic Nerve Segmentation Database http://www.ia.uned.es/~ejcarmona/DRIONS-DB.html

[33] M. Ortega, M. G. Penedo, J. Rouco, N. Barreira, M. J. Carreira. 2009. Retinal verification using a feature points based biometric pattern. EURASIP Journal on Advances in Signal Processing. 2009(Article ID 235746): 13.

[34] VARIA database of retinal images, http://www.varpa. es/varia.html

[35] IIT Delhi Ear Database, http://www4.comp.polyu.edu.hk /~csajaykr/ IITD/Database _Ear.htm

[36] Ajay Kumar and Chenye Wu. 2012. Automated human identification using ear imaging. Pattern Recognition. 41(5).

[37] Hoover A., Kouznetsova V., Goldbaum M. 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical Imaging. 19(3): 203-210.