International Journal of Engineering and Information Systems (IJEAIS)
ISSN: 2000-000X
Vol. 1 Issue 9, November – 2017, Pages: 123-128

Big-data Visualization Platform (Data-mining Approach) for Human Identity Verification

[1]Emdad, S.M. Hossain, [2]Girija Chetty
[1,] University of Nizwa, emdad.hossain@unizwa.edu.om
[2.] University of Canberra, girija.chetty@canberra.edu.au

Abstract: *Verification of human identity is becoming a challenging task due to advance manipulation technique. Likewise, to be on the right stream of identity verification; researcher has been placing their continuous effort. In this paper we are going to present a new and advance technique for human identity verification. For our experiment we used 150 different video sequences for 50 different people. For dimensionality reduction and feature extraction we used linear discriminant analysis (LDA). To get the verified identity we used extracted feature in big-data visualization model using Learning Classifier "Bagging", and our experimental result shows the significance of big-data visualization approach in human identity verification.*

Keywords: manipulation technique, identity, verification, LDA.

1. INTRODUCTION

Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be applied to analyze and extract patterns from large-scale data. The rise of Big Data has been caused by increased data storage capabilities, increased computational processing power, and availability of increased volumes of data, which give organization more data than they have computing resources and technologies to process. We note that the aim of this paper is not to extensively cover Big Data, but present a brief overview of its key concepts with some potential applicability [1] . In this paper we are going to use high dimensional video data to establish different pattern. Perhaps; we are using linear discriminant analysis (LDA) to reduce the dimensionality. And then we will use big-data visualization platform in order to establish the person/human identity.

2. BACKGROUND

Biometric person identification is a common technological tool for identity verification. It carries significant importance for national or international security. All most each and every part of human body is unique; some of the significant ones have been used for developing automate identity verification systems. Fingerprint, palm print, face, iris, ear etc. have been used immensely for current generation of person authentication technologies [2]. Among those entire biometric traits, face recognition still very famous on regards to its effectiveness and convenience to deployment. Research shows in terms of universality, collectability and acceptability face-biometric still keeping its position in number-one [3]. Even though it has number of drawbacks; only by considering all stated point above researcher keeps on moving with face-biometric and doing all sort of cause-and-effect analysis with the ambition of accurate identity establishment [3]. In fact, accurate recognition using face-biometric may be difficult with glasses, masks, long hair etc. we also must ask users to have a neutral face when pictures are being taken. Furthermore, Variations between the images of the same face due to illumination and viewing direction are almost always larger than image variations due to changes in face identity [4]. These variations are exacerbated by additional factors, such as facial expression, perspiration, hair style, cosmetics, and even changes due to aging. After long journey with all stated difficulties above; researcher came up with number of solutions, example; in the last decade, appearance-based methods have been introduced that use low-dimensional representations of images of objects or faces [5].
Appearance-based methods differ from feature-based techniques in that low-dimensional representations are faithful to the original image in a least-squares sense. Techniques such as SLAM [6] have demonstrated that appearance-based methods are both accurate and easy to use. The feature vector used for classification in these systems is a linear projection of the face image in a lower dimensional linear subspace. In extreme cases, the feature vector is chosen as the entire images, with each element in the feature vector taken from a pixel in the image. Despite their success, many appearance-based methods have a serious drawback. Recognition of a face under particular lighting conditions, in a particular pose, and with a particular expression is reliable only if the face has been previously seen

under similar circumstances. In fact, variations in appearance between images of the same person confound appearance-based methods.

In an effort to overcome this shortcoming, there has been a recent surge in work on 3-D face recognition. The idea of these systems is to build face-recognition systems that use a handful of images acquired at enrollment time to estimate models of the 3-D shape of each face. The 3-D models can then be used to render images of each face synthetically in novel poses and lighting conditions—effectively expanding the gallery set for each face. Alternatively, 3-D models can be used in an iterative fitting process in which the model for each face is rotated, aligned, and synthetically illuminated to match the probe image. Conversely, the models can be used to warp a probe image of a face back to a canonical frontal point of view and lighting condition. In both of these cases, the identity chosen corresponds to the model with the best fit [7].

The 3-D models of the face shape can be estimated by a variety of methods. In the simplest methods, the face shape is assumed to be a generic average of a large collection of sample face shapes acquired from laser range scans. Georghiades and colleagues [8] estimated the face shape from changes in the shading in multiple enrollment images of the same face under varying lighting conditions. Although recent advances in 3-D face recognition have gone a long way toward addressing the complications causes by changes in pose and lighting, a great deal remains to be done. Natural outdoor lighting makes face recognition difficult, not simply because of the strong shadows cast by a light source such as the sun, but also because of subjects tend to distort their faces when illuminated by a strong light.

By considering all possible threats on accurate identification; researcher now-a-days moving forward to search for robust approach, robust technique e.g. different methods, classifiers, algorithms etc. in this paper we are proposing a big-data visualization model to have accurate identification of a human. Potentially our proposed model will be effective for any biometric trait like; face, ear, fingerprint, gait etc.

3. METHODOLOGY

In order to have successful completion with potential result; we selected our methodology by considering two important points keeping in mind that are timing and accuracy. We selected data collection, data analysis, dimensionality reduction and experiment methodology for maximum accuracy with minimum time. For our data collection method, we applied following rules [9] which is part of Hong Kong national biometric data collection rules:

1. Need for caution to handle sensitive biometric data
2. Justifications for collecting and using biometric data
3. Risk minimization techniques in biometric data collection
4. The need for a privacy impact assessment
5. Free and informed choice to allow collection of one's biometric data
6. Privacy requirements for dealing with the biometric data collected

After successful collection of our data, we used virtual dub as analytical tools followed by image analysis using imagJ tools. As soon as every single image got ready; we reduced dimensions of those data by using Linear Discriminant Analysis (LDA). However, Linear Discriminant Analysis (LDA) is most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications.

The goal is to project a dataset onto a lower-dimensional space with good class-separability in order avoid over-fitting ("curse of dimensionality") and also reduce computational costs. Ronald A. Fisher formulated the Linear Discriminant in 1936 (The Use of Multiple Measurements in Taxonomic Problems), and it also has some practical uses as classifier. The original Linear discriminant was described for a 2-class problem, and it was then later generalized as "multi-class Linear Discriminant Analysis" or "Multiple Discriminant Analysis" by C. R. Rao in 1948 (The utilization of multiple measurements in problems of biological classification). The general LDA approach is very similar to a Principal Component Analysis.

But in addition to finding the component axes that maximize the variance of our data (PCA), we are additionally interested in the axes that maximize the separation between multiple classes (LDA). Therefore, in a nutshell, often the goal of an LDA is to project a feature space (a dataset n-dimensional samples) onto a smaller subspace k (where $k \leq n-1$) while maintaining the class-discriminatory information. In general, dimensionality reduction does not only help reducing computational costs for a given classification task, but it can also be helpful to avoid over-fitting by minimizing the error in parameter estimation ("curse of dimensionality") [10]; that is one of main reason why we used LDA instead of PCA.

Let's explain how it works for us. Assume **w**1,**w**2, . . . ,**w**c and $n1$, $n2$, . . . , nc denote the classes and the number of concatenated feature vectors h within each class, respectively, with **w** = **w**1 ∪ **w**2 ∪ · · · ∪ **w**c and n^ = $n1$ + $n2$ +· · ·+nc. Note that the value of n^ is two times of n. c is the number of classes. LDA seeks a transformation matrix **W** that maximizes the ratio of the between-class scatter matrix S_B to the within-class scatter matrix

$\sum_{i=1}^{c} ni(Mi-M)(Mi-M)^T$ ....……………………….(1)

$$S_w = J(w)|WTS_BW| \Big/ |WTS_wW|$$

The within-class scatter matrix is

$$S_W = \sum_{i=1}^{c} \sum_{h \varepsilon w_i} (h - Mi)(h - Mi)^T$$

and the between-class scatter matrix is

$$\boldsymbol{S_B} = \sum_{i=1}^{c} ni(Mi - M)(Mi - M)^T \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where $M_i = M_i = (1/ni)\sum_{h \varepsilon w_i}$

and $M = (1/n^{\wedge})\sum_{h \varepsilon w} h$ are the means of the class i and the grand mean, respectively.

We tried to generate the maximum number of concatenated feature vectors based on the characteristics of face [11]. From the extracted feature vector we identified people using big-data visualization and classification approach. Following figure (figure 1) shows how we identified people using big-data visualization/data-mining approach. From the extracted Eigen value; to identify a person we used a learning classifier called bagging. The bagging classifier is a ``bootstrap"[12] and ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set [13] . Research shows that Bagging is effective on ``unstable" learning algorithms where small changes in the training set result in large changes in predictions [12].

It also reduces variance and helps to avoid over fitting, and that is the main significance of the classifier in order to implement into our project. Perhaps; the formula works as: y ~ x + w + z, where y is the response and x, w, z are predictors [14]. Principally it works like kNN classifier, but it is superior in higher dimensions. Bagging in pattern recognition is not new, but the combination of LDA-Bagging provided promising output in our experiment.

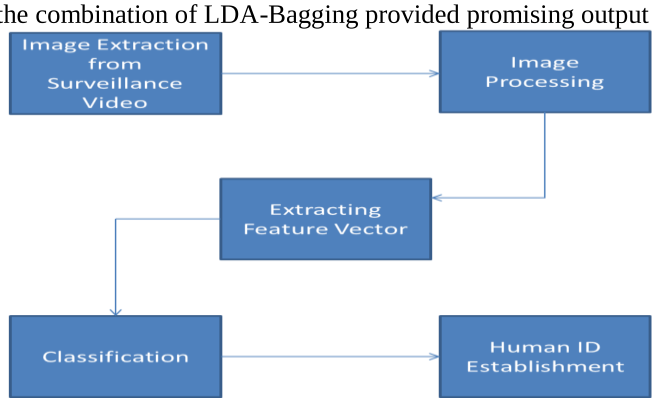

Figure 1: Work-flow of the experiments

4. EXPERIMENTS AND RESULTS

Since we declare earlier in the background of this paper; our proposed model will be effective for any biometric trait like; face, ear, fingerprint, gait etc. Therefore, in our experimental evaluation we applied multi-mode gait fusion scheme, and we used data from CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences [15]. It is a large multi-view gait database, which is created in January 2005. There are more than 300 subjects. We used three (3) different datasets known as dataset A (Normal walking) dataset B (Fast walking) and Dataset C (Walking with bag). All data was captured with infrared camera. The videos were captured at night. For all the experiments, we used 50 subjects from each of the dataset. It means, we used 50 subjects of extracted silhouettes from Dataset A, 50 subjects from B and 50 subjects from C. Each subject consists 16 images and in total 2400 images for 150 subjects.

However, the experiments involved a training phase and a test phase. We used a 10-fold cross-validation for dividing the complete data-form into training and test subsets. With 10-fold cross-validation, the original dataset is randomly partitioned into 10 subsets of the k subsets. A single subset was retained as the validation data for testing the model, and the remaining 9 subsets were used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsets used exactly once as the validation data. The 10 results from the folds were then averaged to produce a single estimation. We found that advantage of this method over repeated random sub-sampling is; all observations could be used for training and validation/testing, and each observation could be used for validation exactly once. In training phase, we built the gait templates for each person using LDA feature vectors for each of the dataset (Dataset A, B and C). In test phase the LDA feature vectors from unseen images in training set were classified with bagging classifier for each of the datasets separately (dataset A, B and C) and by fusion of multiple-mode of walking. Table 1 shows the rate of true identification in the set of experiment

Table 1: Identification in LDA-Bagging protocol

No	Protocol	Dataset	Accuracy
1	LDA-Bagging	Normal Walking	**82%**
3	LDA-Bagging	Fast Walking	**84.88%**
4	LDA-Bagging	Walking with Bag	**88.12%**
5	LDA-Bagging	Fusion of Normal Walking and Fast Walking	**91.38%**
6	**LDA-Bagging**	**Fusion of Normal Walking and Walking with Bag**	**90.25%**

The result shows high level of accuracy with the proposed scheme, for dataset of "walking with bag". We achieved 88.12% correct identification by using LDA-Bagging approach. And only 11.88% has been identified with wrong/incorrect identification. On the other hand, the dataset of "fast walking and normal walking" resulted with 85% and 82% respectively. This is quite interesting to see difference in between the mentioned dataset. In fact, now we are very much clear on shortcomings of identification using gait profile. The result may vary on walking condition. But, most important thing is; it is still identifiable in any walking condition/style.

However, after individual testing of each dataset, we move to fusion of "normal walking and fast walking" and fusion of "normal walking and walking with school bag". The fusion result shows, fusion of normal walking and fast walking giving us better result in compare to fusion of "normal walking" and "walking with bag". Still the difference is not much (91.38% and 90.25%). Figure 2 shows the graphical presentation of achieved result

International Journal of Engineering and Information Systems (IJEAIS)
ISSN: 2000-000X
Vol. 1 Issue 9, November – 2017, Pages: 123-128

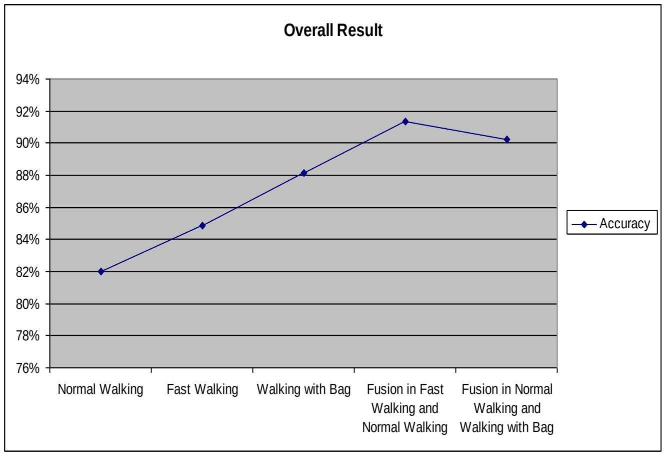

Figure 2: Identification in LDA-Bagging protocol

Moreover, the fusion over individual dataset seems promising, and identification accuracy of each class individually was also good with excellent true positive (TP) rates. The figure 3 shows the comparison of first 20 class of our experiment.

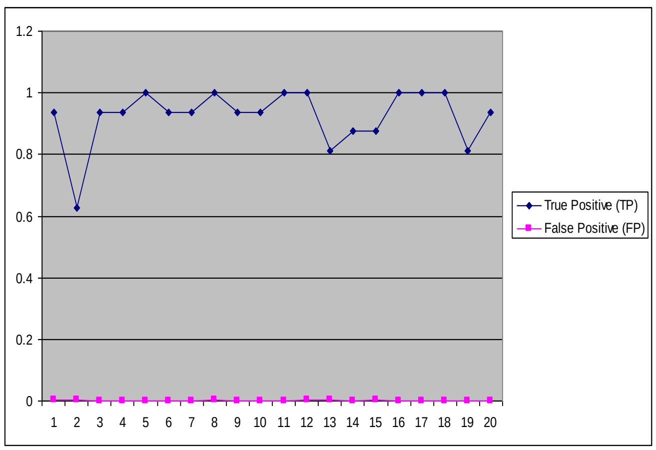

Figure 3: Identification accuracy by class

As we have written early in this paper, in total we had 50 subjects (person) with 3 different dataset (50*3). Each subject represents a class in our experiment. The figure above clearly indicating that the true positive (TP) rates are superior over false positive (FP) rate. FP rate seems in the ground in-compare to TP. However, in summary of our experiment we can say, by using multi-mode of walking in surveillance video footage with long range videos (without detailed face images), it is possible to perform large scale identification with high level of accuracy, using simple subspace features (LDA) and simple data-mining classification technique (bagging). Such simple approaches can lead to real time and real world intelligent video surveillance systems - the beginning of a new dimension of security systems in public surveillance. Our small experimental efforts reported here shows the importance of application of Data-mining approach in surveillance security. And our promising results providing clear indication that it can be applicable to all most all kinds of biometric trait e.g. face, gait, ear fingerprint etc. perhaps; we proved. At the end, we strongly hope and expect; our small experiment will lead us to further experiment in establishment of data-mining technique in person identification model.

5. Conclusions

We can summarize our finding into two parts. One is; identifying a person using their walking pattern/style even though they run, carrying heavy bag, using mask or use different cloths. Another important finding is; using data-mining technique on establishment of human identity. It was completely new to us to apply data-mining methodology in person identification. After our first and successful experiment; we were planning to move forward with it; and we did. Perhaps, our experimental results show that it is an up-coming and promising approach on-establishment of human identity with minimum error and we are very much confident on it.

REFERENCES

[1] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, “Deep learning applications and challenges in big data analytics,” J. Big Data, vol. 2, p. 1, Feb. 2015.

[2] S. M. E. Hossain and G. Chetty, “Person Identity Verification Based on Multimodal Face-Gait Fusion,” IJCSNS, vol. 11, no. 6, pp. 77–86, 2011.

[3] P. T. K, “A Comparative Study of Biometric Technologies with Reference to Human Interface,” Int. J. Comput. Appl., vol. 14, no. 5, pp. 10–15, Dec. 2011.

[4] “Expressions Invarient Face Recognition (PDF Download Available),” ResearchGate. [Online]. Available: https://www.researchgate.net/publication/282878757_Expressions_Invarient_Face_Recognition. [Accessed: 22-Aug-2017].

[5] C. H. Liu and J. Ward, “Face Recognition in Pictures is Affected by Perspective Transformation but Not by the Centre of Projection,” Perception, vol. 35, no. 12, pp. 1637–1650, Dec. 2006.

[6] J. Mundy et al., “An experimental comparison of appearance and geometric model based recognition,” in Object Representation in Computer Vision II, 1996, pp. 247–269.

[7] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, “Eigenfaces vs. Fisherfaces: recognition using class specific linear projection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 711–720, 1997.

[8] M. Singh and A. S. Arora, “Varying Illumination and Pose Conditions in Face Recognition,” Procedia Comput. Sci., vol. 85, pp. 691–695, Jan. 2016.

[9] Guidence Notice, “Guidance on Collection and Use of Biometric Data.” Office of the privacy commissioner for personal data, Hong Kong, Jul-2015.

[10] “Linear Discriminant Analysis,” Sebastian Raschka’s Website, 03-Aug-2014. [Online]. Available: sebastianraschka.com/Articles/2014_python_lda.html. [Accessed: 22-Aug-2017].

[11] E. Hossain and G. Chetty, “Multimodal Feature Learning For Gait Biometric Based Human Identity Recognition,” in 20th International Conference on Neural Information Processing, Gaegu, South Korea, 2013, vol. 20.

[12] L. Breiman, “Bagging Predictors,” Mach. Learn., vol. 24, no. 2, pp. 123–140, Aug. 1996.

[13] B. Efron and R. Tibshirani, An Introduction to the Bootstrap. Chapman & Hall/CRC, 1993.

[14] “bagging function | R Documentation.” [Online]. Available: https://www.rdocumentation.org/packages/ipred/versions/0.4-0/topics/bagging. [Accessed: 28-Aug-2017].

[15] S. Zheng, “CASIA Gait Database,” Inst. Autom. Chin. Acad. Sci. Wwwsinobiometricscom.