Multi-view Gait Fusion for Large Scale Human Identification in Surveillance Videos

Emdad Hossain and Girija Chetty

Faculty of Information Science and Engineering, University of Canberra
emdad.hossain@canberra.edu.au

Abstract. In this paper we propose a novel multi-view feature fusion of gait biometric information in surveillance videos for large scale human identification. The experimental evaluation on low resolution surveillance video images from a publicly available database [1] showed that the combined LDA-MLP technique turns out to be a powerful method for capturing identity specific information from walking gait patterns. The multi-view fusion at feature level allows complementarity of multiple camera views in surveillance scenarios to be exploited for improvement of identity recognition performance.

Keywords: multi view images, LDA, MLP, identification, feature fusion.

1 Introduction

Human identification from arbitrary views is a very challenging problem, especially when one is walking at a distance. Over the last few years, recognizing identity from gait patterns has become a popular area of research in biometrics and computer vision, and one of the most successful applications of image analysis and understanding. Also, gait recognition is being considered as a next-generation recognition technology, with applicability to many civilian and high security environments such as airports, banks, military bases, car parks, railway stations etc. For these application scenarios, it is not possible to capture the frontal face, and is of low resolution. Hence most of traditional approaches used for face recognition fail; however, several studies have shown that it is possible to identify human from a distance from their gait or the way they walk. Even if frontal face is not visible, it is possible to establish the identity of the person using certain static and dynamic cues such as from face, ear, walking style, hand motion during walking etc. If automatic identification systems can be built based on this concept, it will be a great contribution to surveillance and security area.

However, each of these cues or traits captured from long range low resolution surveillance videos on its own are not powerful enough for ascertaining identity, A combination or fusion of each of them, along with an automatic processing technique can result in satisfactory recognition accuracies. In this paper, we propose usage of full profile silhouettes of persons without frontal faces acquired from multiple views, for capturing complementarity or inherent multi-modality available from the gait patterns of the walking human. This also addresses the problems with frontal faces, such as vulnerability to pose, illumination and expression variations. In addition, one of the

J. Blanc-Talon et al. (Eds.): ACIVS 2012, LNCS 7517, pp. 527–537, 2012.

biggest shortcomings of frontal face is user cooperation - a mandatory requirement for establishing identity. On other hand, long range biometric information from surveillance videos captures several biometric traits such as side face, ear, body shape, and gait, which are a combination of physiological and behavioral biometrics, and this rich complementary information can be used in development of robust identification approaches. Further, by using certain automatic processing techniques for extracting salient features based on subspace or kernel methods, multivariate statistical techniques and learning classifiers, it is possible to enhance the performance in real world operating scenarios. In this paper we propose use of complementary information available from multiple views, and simple feature extraction technique based on linear discriminant analysis (LDA) along with a learning classifier based on "MultiLayerPerceptron" (MLP) for establishing identity. Further, we propose a feature level fusion of multiple views as fusing information at an early stage, is more effective than at later stages (score level fusion or late fusion), because features extracted from different biometrics at feature level can retain inherent multimodality much better at feature level and much more information than those in other fusion stages [2]. The experimental evaluation of the proposed multi-view fusion scheme on a publicly available (CASIA [1]) database shows promising performance for real world video surveillance scenarios. Rest of the paper is organised as follows. The background on the role of gait biometric for establishing identity is described in next section. The details of the proposed multimodal identification scheme is described in Section 3. Section 4 describes the experimental setup and results, and Section 5 concludes the paper with some plans for further research.

2 Background

Current state-of-the-art video surveillance systems, when used for recognizing the identity of the person in the scene, cannot perform very well due to low quality video or inappropriate processing techniques. Though much progress has been made in the past decade on visual based automatic person identification through utilizing different biometrics, including face recognition, iris and fingerprint recognition, each of these techniques work satisfactorily in highly controlled operating environments such as border control or immigration check points, under constrained illumination, pose and facial expression variations. To address the next generation security and surveillance requirements for not just high security environments, but also day-to-day civilian access control applications, we need a robust and invariant biometric trait [3] to identify a person for both controlled and uncontrolled operational environments. In this case, trait selection can play vital role. According to authors in [4], the expectations of next generation identity verification involve addressing issues related to application requirements, user concern and integration. Some of the suggestions made to address these issues were use of non-intrusive biometric traits, role of soft biometrics or dominant primary and non-dominant secondary identifiers and importance of novel automatic processing techniques. To conform to these recommendations; often there is a need to combine multiple physiological and behavioral biometric cues, leading to so called multimodal biometric identification system.

Each of the traits, physiological or behavioral have distinct advantages, for example; the behavioral biometrics can be collected non-obtrusively or even without the knowledge of the user. Behavioral data often does not require any special hardware (other than low cost off the shelf surveillance camera), so, it is very much cost effective. While most behavioral biometrics is not unique enough to provide reliable human identification they have been proved to be sufficiently accurate [5]. Gait, is such a powerful behavioral biometric, but on its own it cannot be considered as a strong biometric to identify a person. But, if we combine some other equally non-intrusive biometric with gait; it is expected to be strong combination for human identification. This could be profile (side) images containing side face or ear biometric traits and used with gait. Here side-face and ear images form the physiological component. Both can be collected unobtrusively without user involvement which is very much important in the public surveillance scenarios. It is possible to capture some or all of these multimodal components, if we use gait image information from multiple camera views, which can capture static and dynamic gait profile of a person from one view, with clear side face and ear from other views. This could be extremely applicable and reliable, as most of security infrastructure in public surveillance scenarios currently use multiple cameras. A multi-modal scheme based on such novel approach using multiple camera views can result in establishing identity from long range video images, which is otherwise difficult because face is not clearly visible in such scenarios. Further, it can also address shortcomings of unimodal biometric systems, which perform person recognition based on a single source of biometric, and are often affected by problems such as noisy sensor data and non-universality. Thus, due to these practical problems, the error rates associated with unimodal biometric systems are quite high and consequently it makes them unacceptable for deployment in security critical applications [6] like public surveillance.

Researchers found that one of the most promising techniques is use of multimodality or combination of biometric traits. Using PCA on combined image of ear and face, researchers in [7, 8] have found that multimodal recognition results in significant improvement over either individual biometrics. But most of these schemes work on highly controlled environment which is not quite the case for real world surveillance scenarios. Recently, few attempts have been expended on combining various biometrics in a bid to improve upon the recognition accuracy of classifiers that are based on a single biometric. Some biometric combinations which have been experimented include face, fingerprint and hand geometry [9]; face, fingerprint and speech [10]; face and iris [11]; face and ear [12]; and face and speech [13]. The multi-view fusion in gait profile however, did not attract much attention from the research community. This could be due to difficulty in processing and shortage of multi-view surveillance data. Next Section presents the proposed multi-view gait fusion scheme.

3 Multiview Gait Fusion Scheme

For experimental evaluation of the proposed multiview gait fusion schems, we used CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences [1]. 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 (36 degree view point) dataset B (90 degree view point) and Dataset C (126 degree view point). All data was captured with normal video camera in 11 different views know as view angles. It takes into account four walking conditions: normal walking, slow walking, fast walking, and normal walking with a bag. All of our data here in this experiment taken from normal walking with free hand. The videos were all captured at night. Figure 1 shows the sample images in different view angles.

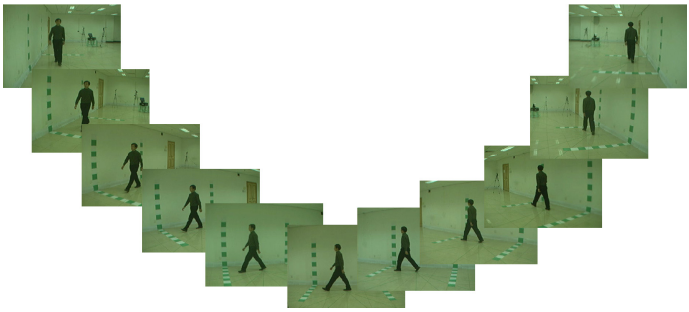

Fig. 1. Sample images

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 of 16 images and in total 2400 images for 150 subjects. Figure 2 shows the extracted silhouettes from dataset B and C.

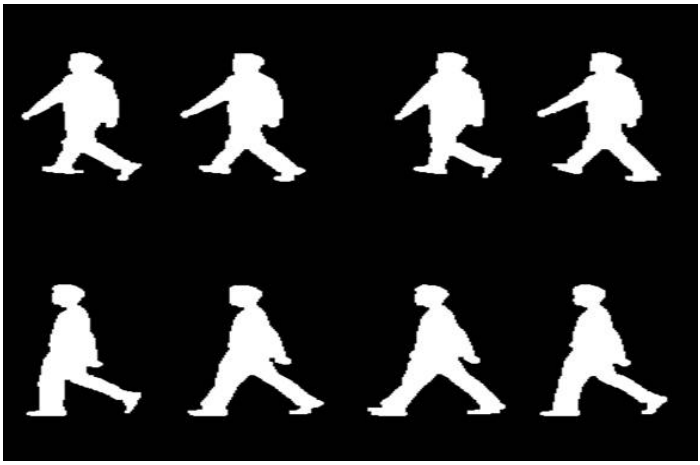

Fig. 2. Extracted silhouettes

For each of the images in these data sets, we extracted the feature vectors in lower dimensional subspaces separately by using PCA (principal component analysis) and Linear Discriminant Analysis (LDA), and used a learning classifier based on well

know multi-layer perceptron (MLP) for classifying each person ID. Our multiview fusion experiments involved identity recognition in LDA-MLP subspace for dataset (unimodal) and fusion of multiple views. The details of LDA subspace for extracting discriminating features is described next.

3.1 Linear Discriminant Analysis

The Linear Discriminent Analysis (LDA) similar to principal component analysis (PCA) and factor analysis, looks for linear combinations of variables which can best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made. LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis [15]. In our experiment LDA shows very promising as LDA model the difference between class and data. Figure 3 shows the extracted feature using LDA

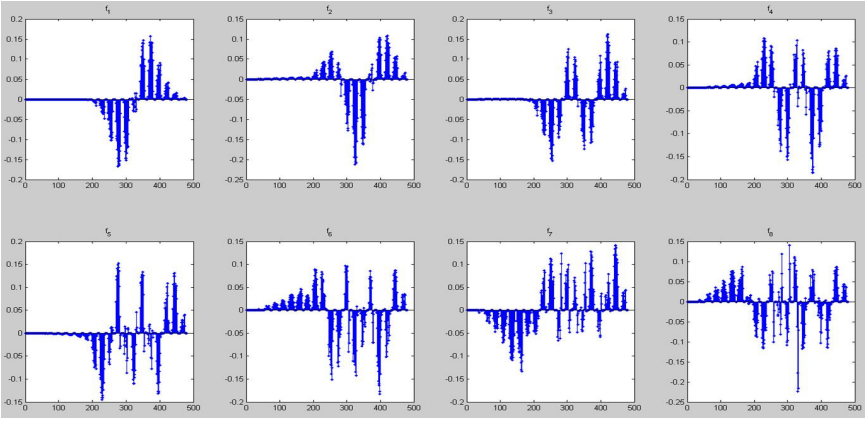

Fig. 3. LDA values extracted from silhouettes

3.2 Multi Layer Perceptron

Multi Layer perceptron (MLP) is a feedforward neural network with one or more layers between input and output layer. Feedforward means that data flows in one direction from input to output layer (forward). This type of network is trained with the backpropagation learning algorithm. MLPs are widely used for pattern classification, recognition, prediction and approximation. Multi Layer Perceptron can solve problems which are not linearly separable 16]. In our experiments we had 49 input

layer, 800 hidden layer (for each data set) and 50 output layer. This is basically based on dimensions, instances and the classes of the dataset. Figure 4 shows the network architecture with MLP, where the green baton to very left; represents dimensions of LDA feature vector, the yellow baton in very right; represents each class (person). The details of the experiments is described in the next Section.

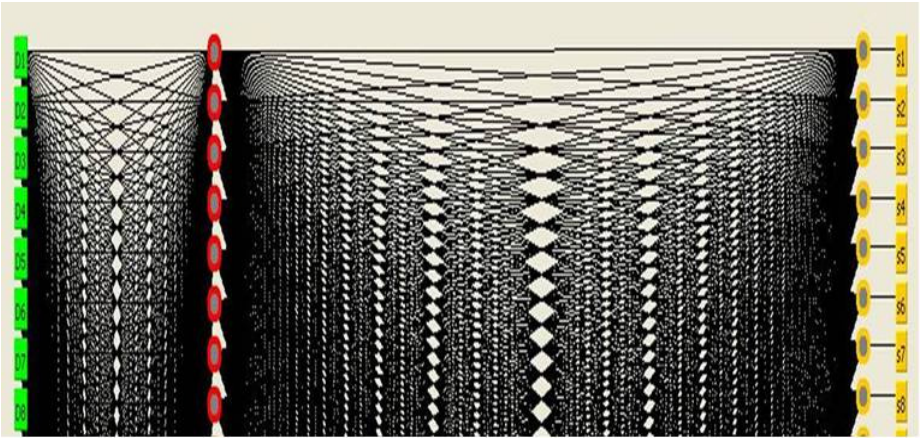

Fig. 4. MLP network architecture for the proposed scheme

4 Experimental Results and Discussion

The experiments involved a training phase and a test phase. We used a 10-fold cross-validation for dividing the complete data from tnto 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 that all observations could be used for both 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 images (Dataset A, B and C) and trained the MLP classifier. In test phase the LDA feature vectors from unseen images in training set were classified with MLP classifier for each of the datasets separately(dataset A, B, C) and by fusion of multiple views. Figure 5 shows the rate of identification in 36 degree view point.

The figure shows high level of accuracy with the proposed scheme, for data captured in 36 degree view point. We achieved 98% correct identification by using LDA-MLP approach. And only 2% has been identified with wrong/incorrect identification. On the other hand, the data captured in 90 degree view point resulted in poor results as compare to the data from 36 degree view point. This could be due to

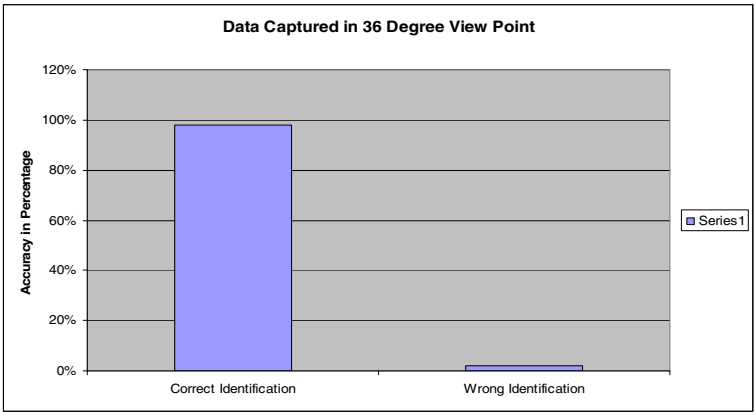

Fig. 5. Rate of Identification in 36 degree view point

difficulty in capturing the identity specific information from 90 degree view point as compared to 36 degrees. Figure 6 shows the results achieved with the data from 90 degree view point.

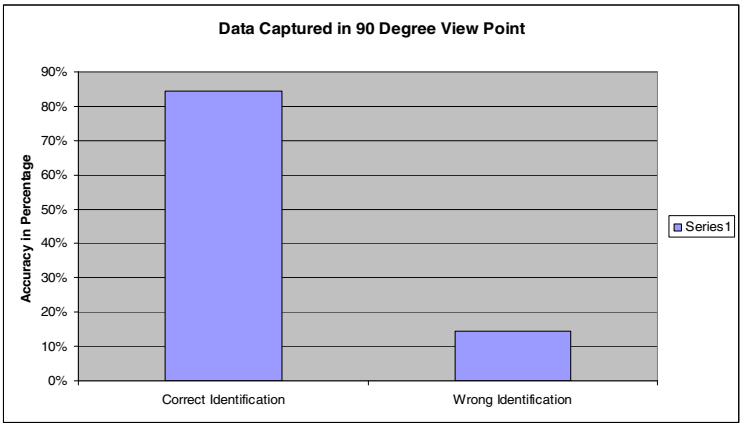

Fig. 6. Rate of Identification in 90 degree view point

The result shows, we received 84.5% correct identification for a large data set which has captured in 90 degree view point. And wrong/incorrect identification rate is around 14.5% which is quite large for real world scenario. Figure 7 represents the identification for dataset C (126 degree view point). It can be seen from this figure that it was possible to achieve 88.88% correct identification with the data captured in 123 degree view point. And 11.12% identified were wrongly identified.

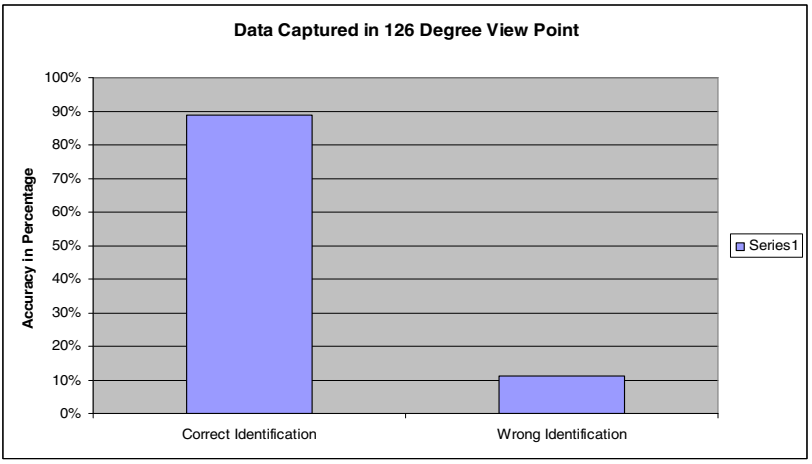

Fig. 7. Rate of Identification in 126 degree view point

After three (3) successful single mode experiments we combined data from all views. We performed feature level fusion of all three extracted set of LDA features, and Figure 8 shows the results of multi-view feature fusion based on gait images from surveillance videos.

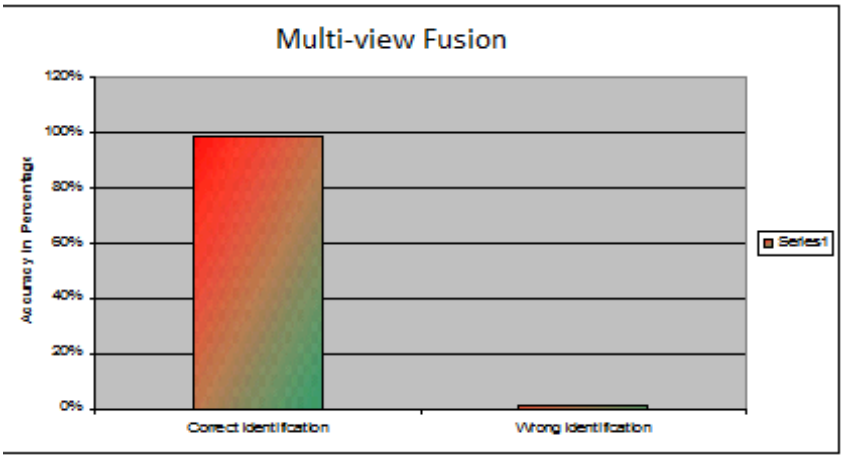

Fig. 8. Result of 3-D fusion

As can be seen from Figure 8, feature level fusion of multiple views results in a significant improvement in correct identification rate as compared to singe views, with 99 % accuracy for fusion of multiple views. Further, accuracy of each class individually was also good with excellent true positive (TP) rates. The figure for details accuracy is shown in the appendix after the references section. As mentioned earlier, each class in the table in the appendix represents each individual or person.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	s1
1	0	1	1	1	1	s2
1	0	1	1	1	1	s3
0.875	0.003	0.875	0.875	0.875	0.994	s4
1	0	1	1	1	1	s5
0.875	0.003	0.875	0.875	0.875	0.998	s6
1	0	1	1	1	1	s7
1	0	1	1	1	1	s8
0.938	0	1	0.938	0.968	1	s9
0.938	0.001	0.938	0.938	0.938	1	s10
0.938	0	1	0.938	0.968	1	s11
1	0	1	1	1	1	s12
1	0	1	1	1	1	s13
1	0.001	0.941	1	0.97	1	s14
1	0	1	1	1	1	s15
1	0	1	1	1	1	s16
1	0.001	0.941	1	0.97	1	s17
1	0	1	1	1	1	s18
1	0	1	1	1	1	s19
0.938	0	1	0.938	0.968	1	s20
1	0	1	1	1	1	s21
1	0	1	1	1	1	s22
1	0	1	1	1	1	s23
1	0	1	1	1	1	s24
1	0	1	1	1	1	s25
1	0.001	0.941	1	0.97	1	s26
1	0	1	1	1	1	s27
1	0	1	1	1	1	s28
1	0.001	0.941	1	0.97	1	s29
0.938	0	1	0.938	0.968	1	s30
1	0.003	0.889	1	0.941	1	s31
1	0	1	1	1	1	s32
0.875	0	1	0.875	0.933	1	s33
1	0	1	1	1	1	s34
0.941	0	1	0.941	0.97	0.951	s35
1	0.001	0.938	1	0.968	0.999	s36
1	0	1	1	1	1	s37
0.938	0	1	0.938	0.968	1	s38
0.938	0	1	0.938	0.968	1	s39
1	0	1	1	1	1	s40
1	0	1	1	1	1	s41
1	0	1	1	1	1	s42
1	0	1	1	1	1	s43
1	0	1	1	1	1	s44
1	0	1	1	1	1	s45
0.938	0	1	0.938	0.968	1	s46
1	0	1	1	1	1	s47
0.938	0.004	0.833	0.938	0.882	0.999	s48
1	0.001	0.941	1	0.97	1	s49
1	0	1	1	1	1	s50

Figure: Detail accuracy by class (person to person)

Finally, to summarize our experimental validation we can say that; by using multiple views of 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 classifier techniques(MLP). 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 multiview images from several cameras and feature level fusion of multiple views as an efficient gait biometric identification.

5 Conclusions and Further Plan

In this paper we proposed a novel multi view feature fusion from low resolution surveillance video for large scale human identification. We applied three (3) different camera views of image data captured with visible cameras. The experimental results shows the multi view fusion approach worked extremely well, indicating the potential of this approach to real time real world public surveillance applications, a truly next generation of surveillance and security systems. Our future research involves investigating novel approaches for exploiting multimodal complementary information available to enhance the performance of human identification for public video surveillance systems.

Acknowledgments. We are very much pleased and thankful to publicly available tools and databases for this paper. We would like to convey our gratitude to Institute of Automation, Chinese Academy of Sciences, for their excellent Database called "CASIA gait database". We also grateful to Machine Learning Group at University of Waikato for their "Weka" machine learning software. This is really massive software especially in machine learning area.

References

1. Zheng, S.: CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences, CASIA Gait Database, http://www.sinobiometrics.com
2. Huang, L.: Person Recognition By Feature Fusion. Dept. of Engineering Technology Metropolitan State College of Denver Denver. IEEE, USA (2011)
3. Bringer, J., Chabanne, H.: Biometric Identification Paradigm Towards Privacy and Confidentiality Protection. In: Nichols, E.R. (ed.) Biometric: Theory, Application and Issues, pp. 123–141 (2011)
4. Jain, A.K.: Next Generation Biometrics, Department of Computer Science & Engineering. Michigan State University, Department of Brain & Cognitive Engineering, Korea University (2009)
5. Yampolskiy, R.V., Govindaraja, V.: Taxonomy of Behavioral Biometrics. Behavioral Biometrics for Human Identification, 1–43 (2010)

6. Meraoumia, A., Chitroub, S., Bouridane, A.: Fusion of Finger-Knuckle-Print and Palmprint for an Efficient Multi-biometric System of Person Recognition. IEEE Communications Society Subject Matter Experts for Publication in the IEEE ICC (2011)
7. Berretti, S., Bimbo, A., Pala, P.: 3D face recognaition using isogeodesic stripes. IEEE Transaction on Pattern Analysis and Machine Intelligence 32(12) (2010)
8. Yuan, L., Mu, Z., Xu, Z.: Using Ear Biometrics for Personal Recognition, School of Information Engineering, Univ. of Science and Technology Beijing, Beijing, 100083 yuanli64@hotmail.com
9. Ross, A., Jain, A.K.: Information fusion in biometrics. Pattern Recognition Letters 24, 2115–2125 (2003)
10. Jain, A.K., Hong, L., Kulkarni, Y.: A multimodal biometric system using fingerprints, face and speech. In: 2nd Int'l Conf. AVBPA, pp. 182–187 (1999)
11. Wang, Y., Tan, T., Jain, A.K.: Combining face and iris biometrics for identity verification. In: Int'l Conf. AVBPA, pp. 805–813 (2003)
12. Chang, K., et al.: Comparison and Combination of Ear and Face Images in Appearance-Based Biometrics. IEEE Trans. PAMI 25, 1160–1165 (2003)
13. Kittler, J., et al.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20, 226–239 (1998)
14. Smith, L.I.: A tutorial on Principal Components Analysis
15. Linear discriminant analysis, Wikipedia, http://www.wikipedia.org
16. MULTI LAYER PERCEPTRON, http://www.neoroph.sourceforge.net
17. Platt, J.C.: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Microsoft Research, Technical Report MSR-TR-98-14, (17) (1998) jplatt@microsoft.com
18. Shlizerman, I.K., Basri, R.: 3D Face Reconstruction from a Single Image Using a Single Reference Face Shape. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(2) (2011)
19. Hossain, E., Chetty, G.: Multimodal Identity Verification Based on Learning Face and Gait Cues. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011, Part III. LNCS, vol. 7064, pp. 1–8. Springer, Heidelberg (2011)
20. Chin, Y.J., Ong, T.S., Teoh, A.B.J., Goh, M.K.O.: Multimodal Biometrics based Bit Extraction Method for Template Security. Faculty of Information Science and Technology, Multimedia University, Malaysia, School of Electrical and Electronic Engineering. Yonsei University, IEEE, Seoul (2011)
21. Multilayer Perceptron Neural Networks, The Multilayer Perceptron Neural Network Model, http://www.dtreg.com