Biometric Identification using Motion History Images of a Speaker's Lip Movements

Alfredo Grunwald de la Cuesta, Jianguo Zhang and Paul Miller
Queen's University Belfast
{agrunwalddelacuesta01,jianguo.zhang, p.miller}@qub.ac.uk

Abstract

This paper describes a new simple, but effective, approach to speaker verification using video sequences of lip movements. We use Motion History Images (MHI) to provide a biometric template of a spoken word for each speaker. Class-dependent correlation filters are then created by a weighted optimization of training MHI samples. Feature extraction is performed by correlating a test MHI against each correlation filter. A Bayesian classifier is deployed for classification. We carry out an extensive performance evaluation of our approach with respect to the number of training samples and different words. Results clearly show the efficacy of our method.

1. Introduction

Current biometric approaches to person verification suffer from the inability of some people to provide a useful biometric template. This is frequently between 1 in 100 and 1 in 1000. In these cases the system must devise a satisfactory way of maintaining security. A potentially low-cost method of identification which can overcome this drawback, is the analysis of a speaker's lip movements. In addition, this approach could also be used with other biometrics, e.g., speaker and face identification systems, to reduce the number of false positives within a large database of enrolled users. Furthermore, lip movements are a behavioral biometric which are particularly difficult to fake by impostors.

To date, a large amount of work has been done in biometrics [1]. Most of the work focuses on a *single mode* of biometrics, e.g. fingerprint, face, human gait, audio, etc. Recently, the trend has been to build robust person identification systems based on a combination of multiple biometric features, i.e. *multimodal* approaches. In this section, we briefly review work closely related to ours. Early work in this area involves the tracking of lips, in which features are usually extracted via color profiles taken around the lip contours. The resulting features are reduced using a principal component analysis and classification is performed by linear discriminant analysis. A significant improvement in speaker verification under noisy conditions was demonstrated by combining lip movement analysis with speech analysis. In [2] an approach of modeling lip movements by Hidden Markov Models (HMMs) is presented. Each lip movement clip is represented by 2D discrete cosine transform (DCT) coefficients of the optical flow vectors within the mouth region. In [4], speech, lip movements and face images are combined to give robust person identification. In this work, DCTs of intensity normalized mouth images were employed to give static features. These were then combined with an HMM to classify the speaker via log-likelihood. It was found that a multimodal approach combining speech, mouth and face features gave significantly improved performance over a single mode approach under trying test conditions.

Though multimodal approaches seem more robust, the effectiveness of such an approach depends heavily on the efficacy of each individual feature. Furthermore, there is still much room for improvement with respect to single mode approaches. Among them, the analysis of lip movements is a relatively novel and robust approach to biometric identification. In this paper, we focus on the single mode approach. In contrast to previous work, we propose to extract the bio-features of the lip movements using motion history images (MHI) [5]. We demonstrate that such a feature has great potential for use in biometric identification, even with a very simple classification method. The MHI is a scalar-valued image where the intensity is a function of how recent the motion is. This temporal template has been used in real-time approaches to human movement classification [5], where recognition is achieved by statically matching moment-based features derived from the MHI. In this paper we investigate the use of MHIs for biometric identification using image

sequences of mouth movements when speaking. In section 2 we describe how MHIs are generated. Section 3 describes the design of class-dependent correlation filters. Section 4 presents the classifier design. Experimental results are presented in section 5. Section 6 gives a conclusion and summary.

2. Movemetrics: Motion History Images

In most successful biometrics systems, the first key step is to extract effective bio-feature representations for the biometric metadata, e.g, iris [7], face [8], fingerprint [9], human gait [10], ear shape [11] etc. Among these, human gait is the biometric which is most similar to that of lip movements, in that both are of a dynamic nature. In the former, human gaits are represented by extracting features from a set of silhouettes [10]. The generation of silhouettes is based on an assumption that human movement is distinct against the background. However, this assumption is not valid for lip movements, since they are against a face background, making them less distinct. Thus the silhouette based approach is not suitable for the description of lip movements. Furthermore, it is very hard to produce reliable lip contours for the images shown in Figure 1(a).

Inspired by the successful work of Bobick and J. W. Davis on human motion recognition [3], we propose to generate biometric features of lip movements using MHIs. A video sequence of a speaking mouth presents a particular movement which contains the information unique to the person of interest. To extract such information, we model the movement using MHIs. The basic idea of MHIs is to model the motion by accumulating intensity changes of pixels. Specifically, they are scalar-valued images, where the intensity is a function of how recent the motion is [3]. This means that the value of each MHI pixel is a function of the temporal history motion at that point.

Figure 1 shows an example of the MHI construction process. The MHI at time t is calculated according to the following equation:

$$H_\tau(i,j,t) = \begin{cases} \tau & \text{if } D(i,j,t)=1 \\ \max\left[0, H_\tau(i,j,t-1)-1\right] & \text{otherwise} \end{cases} \quad (1)$$

where:

$$D(i,j,t) = I((i,j,t) - I(i,j,t-1) \quad (2)$$

and where $I(i,j,t)$ is a frame at time t of the sequence, Fig. 1(a), and $D(i,j,t)$ is the difference between two consecutive frames, which is normalised to achieve illumination robustness [3]. $D(i,j,t)$ is further binarised by setting a threshold at some value δ between zero and one, as shown in Figure 1(b). The variable τ represents the duration of the speech segment we are interested in, i.e. the number of frames in the sequence.

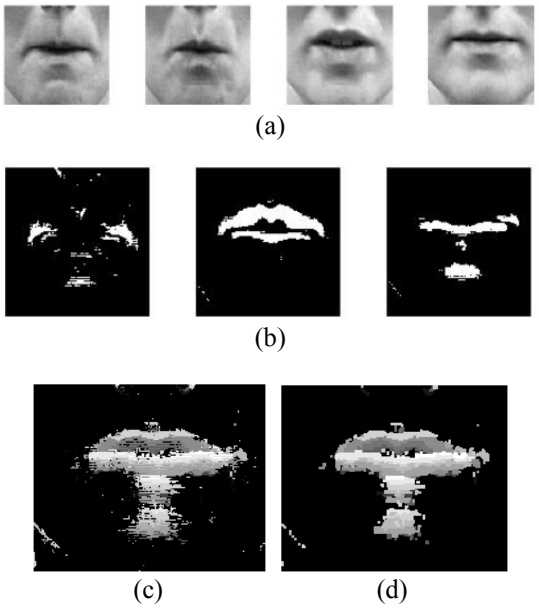

Figure 1. Construction of MHIs. (a) Sequence images of a speaker pronouncing one word. (b) Differencing of consecutives frames, followed by δ=0.4. (c) MHI of the sequence. (d) Noise reduced MHI.

Note that the MHI, Figure 1(c), still contains some noise, which we can reduce by morphological operations. Here we employ opening, which is performed over every $D(i,j,t)$ image after setting the threshold δ. Thus, a perceptible improvement on the MHI is achieved as shown in Figure 1(d).

3. Class-dependent Filters

MHIs can be considered as temporal templates, images where each pixel is a function of the motion at that pixel location. These templates are used to create correlation filters for each model of movement. In this work we use *phase-only* filters, since in matched filtering the phase information is more important than amplitude information, resulting in a reasonably sharp correlation peak [6]. The correlation function is given by:

$$C(m,n) = \Im^{-1}\{T(u,v)F(u,v)\} \quad (3)$$

where $T(u,v) = \Im\{t(i,j)\}$, $F(u,v) = \Im\{f(i,j)\}$, $t(i,j)$ is the MHI of a lip sequence, $f(i,j)$ is the filter impulse response and $\Im$ and $\Im^{-1}$ denote the Fourier transform and its inverse respectively. Conventionally, Eq. (3) corresponds to an autocorrelation when $F(u,v)=T^*(u,v)$,

where * denotes the complex conjugate. The function $F(u,v)$ is in general complex with:

$$F(u,v) = |F(u,v)| \exp\left[i\phi_F(u,v)\right] \quad (4)$$

where $|F(u,v)|$ and $\exp[i\phi_F(u,v)]$ denote the magnitude and phase respectively. We define the phase-only filter as:

$$F(u,v) = \exp\left[i\phi_F(u,v)\right] \quad (5)$$

For the case where there are more than one training sequences, say $t_1(i,j), t_2(i,j), \ldots, t_N(i,j)$, we need to design a filter such that the response from each of the in-class training samples is uniform, i.e. $C_1 = C_2 = \cdots = C_N$. Due to the nonlinearity of the phase-only operation, a closed form analytical solution is not possible. Therefore, we employ a technique in which, firstly, a composite filter is created by:

$$F(u,v) = \frac{\sum_{i=1}^{N} a_i T_i^*(u,v)}{\left|\sum_{i=1}^{N} a_i T_i^*(u,v)\right|} \quad (6)$$

where a_i are a set of weights that are initialised as being equal to C_i/C_1. A gradient descent algorithm is then used to search the a_i space for a set of weights that satisfy a uniformity criterion for the filter response over the whole training MHI image set.

A set of *class-dependent* phase-only filters are created using one or more MHIs from the same class, one filter per class. When an unclassified MHI is filtered, the 2D correlation between this template and those used to create the filter is being performed. As a result of this operation, depending on the correspondence of the class, we obtain a high or low correlation peak value.

In the following example, the result of creating a filter from the MHI of Figure 1(d) and performing the autocorrelation, is shown in Figure 2(a). The correlation obtained with another MHI of the same class is shown in Figure 2(b), giving us an example of matching. Figure 2(c) shows the result when the sample is from another class, hence, there is no matching. When matching occurs, the peak value is appreciably higher, which indicates that the filter and samples agree with the same class label.

4. Classification

Our classification strategy can be formulated in a Bayesian framework. Suppose that we have L person

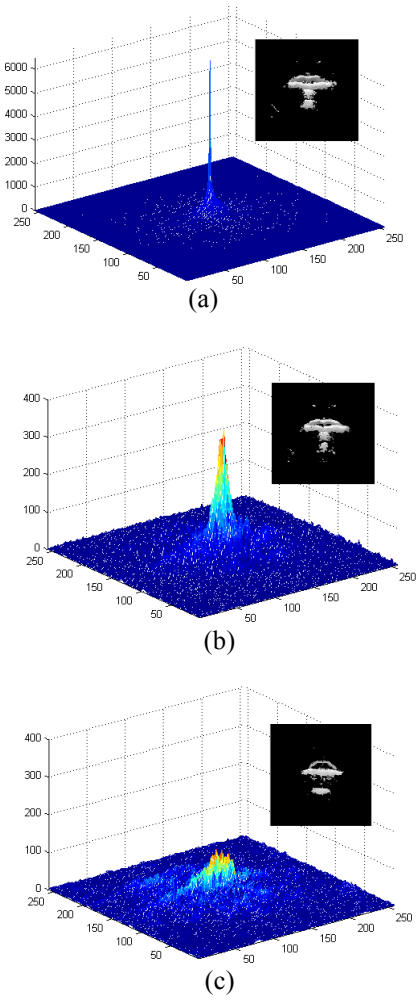

Figure 2. (a) Autocorrelation of an MHI. (b) Correlation of MHIs of the same class. (c) correlation of MHIs of different classes.

classes, $1, \ldots, L$, denoted by y. Given a test example x, we want to compute a posterior $p(y|x)$ as follows:

$$p(y=l|x) = \frac{p(x,y=l)}{p(x)} = \frac{p(x,y=l)}{\sum_{l \in L} P(x,y=l)} \quad (7)$$

For our model, we define the joint probability in terms of the potential function $\Phi(x,y)$:

$$p(x,y=l) = \frac{1}{Z}\Phi(x,y=l) \quad (8)$$

where Z is the partition function with $Z = \sum_{l,x} \Phi(x,y=l)$. Thus, Eq. (8) can be written as:

$$p(y=l|x) = \frac{\Phi(x,y=l)}{\sum_{l} \Phi(x,y=l)} \quad (9)$$

Since y is a discrete variable, it is very easy to perform the summation in denominator. In our approach, we consider the maximum score of the cross correlation, as specified in section 3, as the potential function, thus we have:

$$\Phi(x, y=l) = \max_{m,n}\{C_{y=l}(m,n)\} \tag{10}$$

where $C_{y=l}(m,n)$ is the correlation plane (Eq. (3)) obtained with the template filter, $F_{y=l}(u,v)$, that we have constructed in section 3 for the class l. To compute $C_{y=l}$, we also need $T(u,v)$ which is the Fourier spectrum of the MHI of a test video sample. The potential function can be equivalently computed via Eq. (3). Thus for a given test sample, its class label is determined as follows:

$$l^* = \arg\max_{l \in L} p(y=l \mid x) \tag{11}$$

Our complete verification system is summarised in Fig. 3.

5. Experiments

To investigate the efficacy of our MHI-based approach for speaker identification we carry out a set of experiments: a) Inter- and intra-class robustness; b) identification versus different number of training samples; c) identification across different words. Our database contains 9 different classes corresponding to 9 different speakers. Each person pronounces 9 different words; "one", "two",…, "nine". For each word per person, we have nine video samples. Thus our database contains 9^3=729 video samples. Each video sample presents the movement of the mouth as the dominating area within the image clip. Figure 1(a) shows frames from a sample sequence in our database.

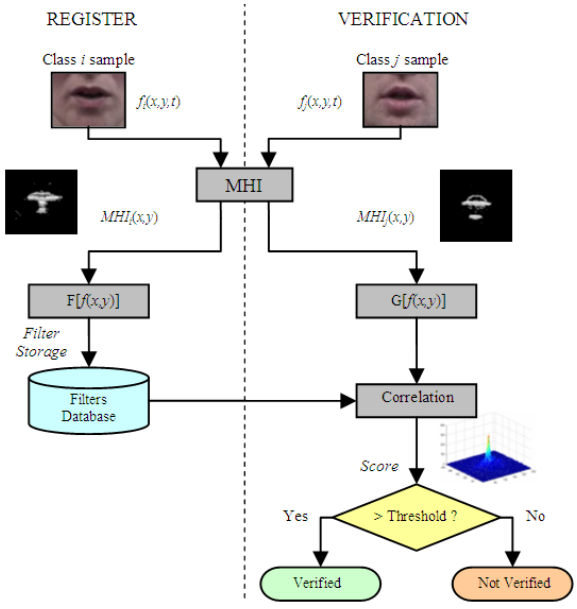

Figure 3. Architecture of the Verification System

5.1 Inter- and Intra-class Robustness

In order to analyze the discrimination power of our features, we calculate the Fisher Ratio:

$$FR = \frac{m_1^2 - m_2^2}{\sigma_1^2 + \sigma_2^2} \tag{12}$$

where m_1 and m_2 are the means of the values of the correlation score respectively; and σ_1^2 and σ_2^2 are the variances of these scores. Table 1 shows the variation in Fisher ratio with MHI threshold.

Word	FR (δ=0.2)	FR (δ=0.4)	FR (δ=0.6)
"ONE"	3.1339	6.4868	2.0081
"TWO"	6.5263	5.9878	3.1344
"FOUR"	5.161	2.9472	2.6859
"EIGHT"	6.9119	6.84	2.9213

Table 1. Averaged Fisher ratio against different thresholds using one training example.

From this table, we can see that the discrimination power with thresholds of 0.2 and 0.4 is much higher than with under 0.6. Figure 4 plots the feature statistics of two classes when the MHI threshold is 0.4. From this figure, we can see the distinctive separation between the two classes, though there is some overlap. This insightful analysis indicates that our feature is discriminative and should work well with a simple classifier. This is further verified by the following experiment.

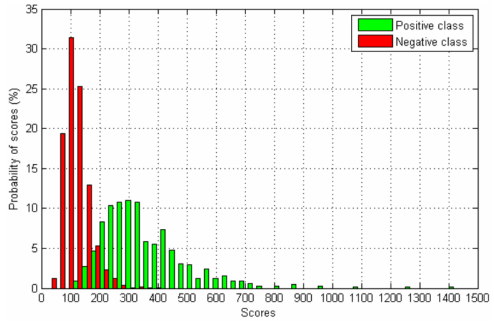

Figure 4. Feature statistics of different classes. Note these are averaged over all words and classes.

5.2. Identification Results vs. Training Samples.

In this experiment, we carry out a quantitative evaluation of the classification performance of our approach. To do this, we perform an experiment to differentiate speakers pronouncing the same word. Instead of producing a single class label for each test sample, we use the *Receiver Operating Characteristic* (ROC) curve as the performance measure. This is a plot of the true positive rate against the false positive rate. The ROC is obtained by varying a threshold corresponding to the classifier output $p(y|x)$. Our experimental details are described as follows.

Filters are created for each class using, at least, one sample of each person. The number of samples used to create our filter varies from 1 to 3. Then, all MHIs of the rest of the samples are filtered. The output of Eq. (9) is considered as the similarity measure.

Considering the word 'one', for each class, we create one set of ROC curves, resulting in 9 sets of different ROC curves. In order to show a more robust measurement, the final ROC curve is the average of those ROC curves. These are shown in Figure 5. We generate three ROC curves with MHI threshold values of 0.2, 0.4 and 0.6.

In our experiment, we consider three possible cases. (1) We first take one sample per class and the rest for testing (9 for training and 72 for testing). In this case the template filter is created by one training sample only. Thus, we have nine different template filters, each one per class. The average ROC curves are shown in Figure 5(a) with 3 possible δ values. (2) In the second case we repeat the experiment by increasing the number of training samples to 2 (18 for training and 63 for testing). The template filters are now constructed by a weighted combination of the two MHI training images according to Eq. (6). Again, 9 filters are constructed. ROC curves are shown in Figure 5(b). (3) The third case is with 3 training samples for each filter (27 for training, 54 for testing). Results are shown in Figure 5(c).

According to these results, it is possible to verify that the best results are obtained when the MHI threshold is set at 0.4 and using three training samples. As shown in Figure 5(c), the recognition rate reaches 100% with a false alarm rate of 5%, which means that this system is quite capable of differentiating speakers saying the same word.

From the results in Figure 5, it is obvious that more training samples produce better results. However, the difference is not significant. With one training sample only per class, our approach can still achieve nearly 95% true positive rate at a level of false alarm rate of 5%, as shown in Figure 5(a). This clearly shows the robustness of our method against the number of training samples. In a real application, this means the user just has to register the keyword once, and then they can be effectively verified in the future.

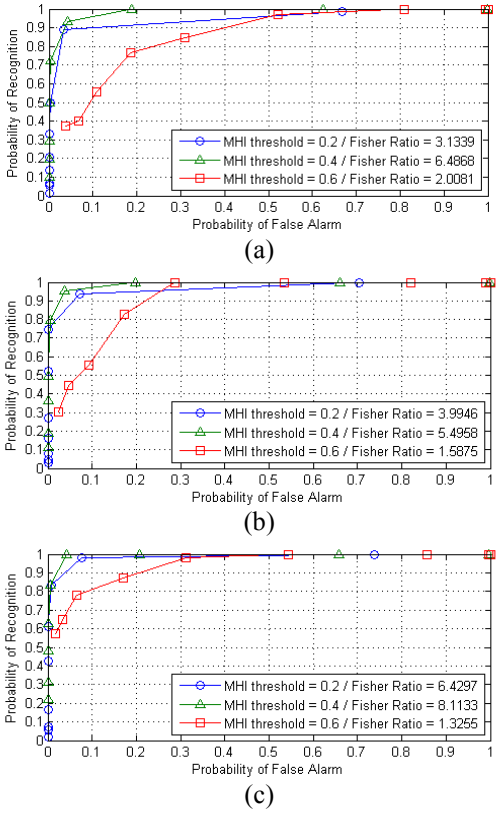

Figure 5. Average ROC curves for all classes against different number of training samples. (a) Using one training sample. (b) Using two training samples. (c) Using three training samples.

5.3 Identification Results vs. Words

In order to test the identification performance of our method against *different* words (word 'one' to 'nine'), an experiment was carried out using one training sample only, i.e., the most difficult case discovered in section 5.2. We plot average ROC curves (averaged over all classes) of each individual word, which means using 81 samples for training and 648 samples for testing. Thus, as we use three different MHI thresholds, we obtain nine sets of ROC curves corresponding to one for each word. Due to space limitations, we only present results for words 'two', 'four', and 'eight', Figure 6(a-c).

From these results, we can see that identification performance does vary against different words.

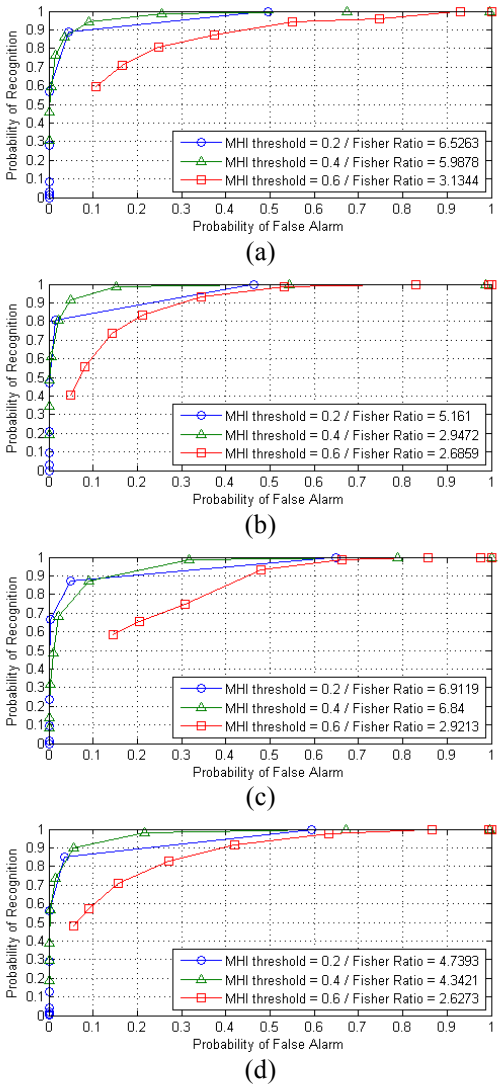

Figure 6. ROC curves for words 'two', (a) 'four', (b), and 'eight', (c). Average ROC curves over all classes and words, (d).

However, the variation is not significant. Our method achieves around 90% recognition rate at the level of 5% false alarm rate for words 'one', 'four' and 'eight'. Figure 6(d) shows the average ROC curve of all of the nine classes and nine words. Again, consistent with the experiment in section 5.2, the best results are obtained with an MHI threshold of 0.4. It is worth noting that our approach still obtains an average recognition rate of 90% at a false alarm rate of 5%. This clearly shows the robustness of our method against different words, i.e., *word independent robustness*.

6. Conclusion

In this paper, we have presented an approach that is capable of identifying humans based on lip movemetrics captured by MHIs. The class similarity is measured using a sample against a set of class-dependent filters, which is created by a weighted optimization of the MHI training samples. We have demonstrated the efficacy of our approach in extensive experiments.

At the moment, our approach models the temporal dependency between consecutive frames as a summation within the MHI images. This is achieved in the representation level. Whether or not it is necessary to specifically learn the dependencies using advanced models in our framework can be investigated in future work. Further work can also employ audio features as an additional cue.

7. References

[1] Anil K. Jain, Sharath Pankanti, Salil Prabhakar, Lin Hong, Arun Ross, "Biometrics: A Grand Challenge," icpr , pp. 935-942, 2004

[2] T. J. Wark, S. Sridharan, and V. Chandran, "The use of speech and lip modalities for robust speaker verification under adverse conditions" in *Proc. Int. Conf. Multimedia Computing and Sustems*, Florence, Italy**,** 1997, pp. 812-816.

[3] Cetingul, H.E. Yemez, Y. Erzin, E. Tekalp, A.M. , **"**Discriminative lip-motion features for biometric speaker identification", in *IEEE ICIP*, 2004, vol.3, pp.2023-2026.

[4] N. A. Fox, R. Gross, J. F. Cohn and R. B. Reilly, "Robust Biometric Person Identification Using Automatic Classifier Fusion of Speech, Mouth and Face Experts", *IEEE Trans. On Multimedia*, **9**, 2007, pp. 701-714.

[5] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates", *IEEE Trans. Pattern Anal. Mach. Intell.*, **23**, 2001, pp. 257-267.

[6] P. C. Miller, "Signal-to-Noise Ratio Analysis for Nonlinear N-ary Phase Filters", Applied Optics, 46, 2007, pp. 6406-6418.

[7] L. Ma, T. Tan, Y. Wang, and D. Zhang. Personal identification based on iris texture analysis. *In IEEE Patt. Anal. Mach. Intell.*, volume 25, pages 1519--1533, 2003.

[8] Jen-Tzung Chien, Chia-Chen Wu, "Discriminant Waveletfaces and Nearest Feature Classifiers for Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* ,vol. 24, no. 12, pp. 1644-1649, December, 2002.

[9] A. K. Jain, L. Hong, and R. Bolle, "On-line Fingerprint Verification," *IEEE Trans. Pattern Anal. and Machine Intell.,* Vol. 19, No. 4, pp. 302-314, 1997

[10] Sarkar, S. Phillips, P.J. Liu, Z. Vega, I.R. Grother, P. Bowyer, K.W. The humanID gait challenge problem: data sets, performance, and analysis In *IEEE Patt. Anal. Mach. Intell.*, volume 27, pages 162--177, 2005

[11] Ping Yan, Kevin W. Bowyer, "Biometric Recognition Using 3D Ear Shape," *IEEE Transactions on Pattern Analysis and Machine Intelligence* ,vol. 29, no. 8, pp. 1297-1308, August, 2007.