Subject Filtering for Passive Biometric Monitoring

Vahan Grigoryan[1], Donald Chiarulli[2], and Milos Hauskrecht[3]

[1] University of Pittsburgh vahan@cs.pitt.edu
[2] University of Pittsburgh don@cs.pitt.edu
[3] University of Pittsburgh milos@cs.pitt.edu

Summary. Biometric data can provide useful information about the person's overall wellness. However, the invasiveness of the data collection process often prevents their wider exploitation. To alleviate this difficulty we are developing a biometric monitoring system that relies on nonintrusive biological traits such as speech and gait. We report on the development of the pattern recognition module of the system that is used to filter out nonsubject data. Our system builds upon a number of signal processing and statistical machine learning techniques to process and filter the data, including, Principal Component Analysis for feature reduction, the Naive Bayes classifier for the gait analysis, and the Mixture of Gaussian classifiers for the voice analysis. The system achieves high accuracy in filtering non-subject data, more specifically, 84% accuracy on the gait channel and 98% accuracy on the voice signal. These results allow us to generate sufficiently accurate data streams for health monitoring purposes

1 Introduction

This research is a part of the "Nursebot Project", which aims to develop a mobile robotic assistant for elderly people who are at risk of institutionalization in order to allow them to maintain independence for as long as possible. The goal of this part of the project is to monitor their wellness, detect any relevant change, and report it to the health care professional. Some people are sensitive about intrusion into their life, therefore, we would like to collect the data inconspicuously. The passive monitoring of wellness will give a better idea of a patient's condition to doctors and nurses.

People have always used biological traits, such as voice, face, gait, etc. to recognize each other. Biometrics emerged as an automated method of identifying individuals or verifying the identity of a person based on distinctive physiological or behavioral characteristics. It is natural to extend biometric analysis systems so that they assess a person's wellness by his/her behavioral and/or physiological traits. The non-intrusion constraint limits our choice of

biometric traits mainly to behavioral ones. Our research develops solutions based on two of them: voice and gait.

Passive biometric monitoring in a real-world environment raises one important problem. Data collected by the sensors will include readings from all individuals who enter and leave the environment and will not be always generated by our target subject. Thus, it is very important to filter the data so that only the target subject is monitored and "biometric noise" from other individuals is rejected. The problem of filtering of non-subject data is related to the problem of machine recognition of human subjects, but there are several important differences. First, we can tolerate a large number of "true-reject" errors. That is because we deal with a long term monitoring of the subject, and low effective sample rate is not an issue for our system. Second, the sensors are placed in the environment in which we expect the presence of a rather limited number of people, e.g. family, friends, and caregivers. That puts an upper bound to a number of people our system detects besides the subject.

The filtering system described in this paper uses data from three sensors: two microphones and an accelerometer. One of the microphones and the accelerometer are used to collect data about the person's gait, and the other microphone is used to monitor vocalizations. The analysis consists of three steps: (1) feature extraction and reduction, (2) learning of discriminatory patterns and (3) filtering. In the first stage, the dimensionality of the data is reduced to a reasonable size that facilitates further analysis. In the learning stage the features are used to extract discriminatory patterns from labeled signal samples. The patterns use information from all three sensors. Filtering of the signal exploits the patterns learned and applies them to the continuous stream of data to identify the target subject.

In the following section we describe the underlying model and methods used in our system. Next we present the results of experiments with filtering of subject data. Finally, we give the summary in Sect. 4.

2 Model Description

2.1 Gait Analysis

Data Acquisition and Feature Extraction

Our system analyzes gait data collected from a piezoelectronic accelerometer and a microphone at sampling rate of 20KHz. The most significant footstep for each pair of "raw" signals is extracted by detecting the largest peak and taking $N = 10000$ data points in its neighborhood. Further processing is done in spectral domain. We take discrete Fourier Transform of the signal, then we use an ideal lowpass filter with a cutoff frequency of 4KHz.

Next we have to reduce our feature dimensionality, since it would be computationally hard to work with 2000 features per sample. One of the

widely used dimensionality reduction methods is Principal Component Analysis (PCA). PCA performs linear transformation that aligns the transformed axis with the directions of maximum variance (cf. Jolliffe, 2002). Principal components are eigenvectors of the covariance matrix of samples, taken in decreasing order of corresponding eigenvalues, i. e. the first principal component is an eigenvector corresponding to the largest eigenvalue of the (sample) covariance matrix. The first few dimensions of PCA transformed data contain most of information about the data. We keep the first 30 features for each of the samples as our main features for classification purposes.

Classification and Fusion

We use the Naive Bayes model to separate the main subject from the rest (cf. Domingos and Pazzani, 1997). We compute the posterior probability of the subject given the feature vector $\mathbf{x}$:

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}, \tag{1}$$

and use it to discriminate between the subject class and the impostor class. The conditional density function $p(\mathbf{x}|\omega_i)$ is modeled by a multivariate normal density. We make the *naive* assumption that features are independent given the class, thus

$$p(\mathbf{x}|\omega_i) = p(x_1|\omega_i)p(x_2|\omega_i)\dots p(x_n|\omega_i). \tag{2}$$

We apply the maximum likelihood principle (cf. Duda et al., 2001) to estimate the parameters of density functions.

The above mentioned data processing has been done for samples from each channel (vibration from accelerometer and audio from microphone) independently from the other channel. The central issue of this research is to combine data collected from different sensors (cf. Brunelli and Falavigna, 1995; Hong et al., 1999) in order to obtain a better model of the patient's condition.

We build a logistic regression model (cf. Duda et al., 2001) for the fusion of vibration and audio data characterizing the gait. The model uses a set of adaptive weights that determine the importance of audio and vibration signals. The results from the Naive Bayes model are supplied as four inputs to a sigmoidal unit. We employ the online gradient decent approach for weight optimization (cf. Haykin, 1999):

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \rho_k(y_k - f(\mathbf{w}_k, \mathbf{a}_k))\mathbf{a}_k, \tag{3}$$

where $\mathbf{w}_k$ is the weight vector, $\mathbf{a}_k$ is the input vector, y_k is the desired output, and ρ is the parameter that scales the gradient update.

2.2 Voice Analysis

Cepstral feature extraction

Speech spectrum is one of the best known characteristics of a speaker (cf. Atal, 1976). Therefore we proceed with frame-based spectral analysis as described

in (cf. Reynolds and Rose, 1995). We build the spectrogram of the speech signal by taking the short-time Fourier Transform with a Hamming window of length 25.6 ms ($N = 512$). Next we apply mel-scaled filterbank (cf. Stevens and Volkmann, 1940) of 29 overlapping triangular filters to the spectogram magnitude. We obtain the cepstral coefficients $c_m(n)$ by taking the discrete cosine transform of the logarithm of the filterbank output. Cosine transform helps to decorrelate the data, and thus reduces its dimensionality. Finally, we form our feature vector from 12 cepstral coefficients. This process is repeated every 12.8 ms, producing about 78 feature vectors per second.

Gaussian Mixture Model

Human speech is a complex audio signal, and due to phonetic diversity it would be extremely hard if not impossible to come up with a simple parametric density model that effectively characterizes the speaker. So it is natural to describe it as a mixture of several density of functions. Gaussian Mixture Model was shown to be quite successful in solving speaker and speech recognition tasks (cf. Reynolds and Rose, 1995; Gopinath, 1998).

The density function of the Gaussian mixture with m components is given by

$$p(\mathbf{x}|\theta) = \sum_{m=1}^{M} w_m f(\mathbf{x}|\boldsymbol{\mu}_m, \Sigma_m), \tag{4}$$

where $\mathbf{x}$ is a feature vector, w_m's denote the *mixing weights* and $f(\mathbf{x}|\boldsymbol{\mu}_m, \Sigma_m)$ are multivariate Gaussians. The parameter θ consists of weights, means and covariances of all component densities.

We use the training data for the target speaker to estimate the mean vectors, weights, and component densities for his/her model. In our model each Gaussian component has a diagonal covariance matrix. We proceed with parameter estimation using the Maximum Likelihood principle. Direct computation of parameters is not possible due to their nonlinearity. Therefore, we estimate the parameters iteratively using the Expectation Maximization (EM) algorithm (cf. Dempster et al., 1977).

Identification is performed by using the Bayes rule and comparing the posterior probabilities of the mixture model for the target person with the mixture model for the rest of the subjects. The details about our experimental data and preliminary results are discussed in the next section.

3 Experimental Results

Gait Analysis

We conducted our gait identification experiments on data collected from 22 people. For each person there are 10 audio and vibration recordings, collected in 10 different sessions. However for the target person there are 50 recording.

Based on signal to noise and classification analysis of window length we have set the length of our window of interest at $N = 10000$. The mean signal to noise ratio averaged over the window of interest is equal 2.49 dB for vibration signal and 5.67 dB for audio signal.

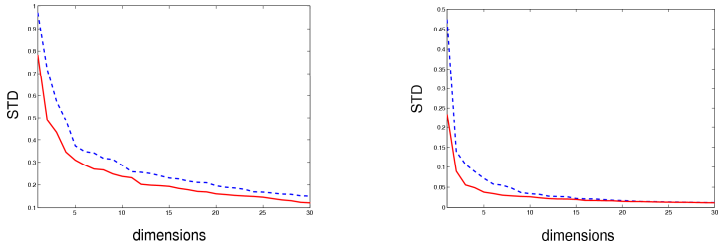

Fig. 1. PCA analysis of impostor and target data. Vibration signal is on the left, audio signal is on the right. The solid line is the standard deviation per dimension for the target subject, the dashed line is the standard deviation per dimension for non-target subjects

The impostor model in out system is based on data that has been determined to be not from the target subject. We have made preliminary statistical analysis of our data in order to justify the choice of our impostor model. We performed PCA on 50 samples from the target subject and on 50 samples from five non-target subjects. As it can be seen from Fig. 1, the PCA transformed data for mixture of subjects displays more variability than the data for a single subject.

Table 1. Averaged confusion matrix of gait recognition by integrated system of audio and video channels

		Actual	
		Target	Impostor
Prediction	Accept	4.44%	1.28%
	Reject	14.79%	79.49%

We could identify the target person by his/her gait with about 84% accuracy, using combined data from audio and vibration channels. The averaged confusion matrix of person recognition by gait based on 30 iterations, and 30%-70% split of data into testing and training sets for each iteration is shown in Table 1.

It is clear from the ROC curves shown in Fig. 2 that by integrating the channels we have improved recognition rates. We can tolerate high false-reject rate, since in our specific case of strong verification problem we are mainly

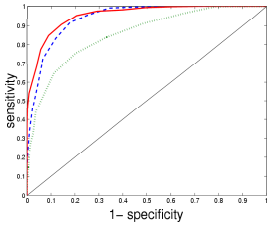

Fig. 2. ROC curves: the sold one is the integrated classifier, the dashed one is the vibration channel only, and the dotted one is the audio channel

interested in low relative false-accept rate, which we define as the ratio of false-accepts to a total number of accepted samples. That is because we are conducting long term monitoring, thus we have access to very large amount of subject data. Therefore we can reject a lot of subject data as long as we keep some (in this case about one-fifth). However we would like to include as few non-subject features in our training set as possible.

The preliminary results show that the relative false-accept rate is 22.38% which is somewhat high. That can be explained by the relatively small amount of testing data, so that in this case only one ($0.0128 * 78 = 0.998$) falsely accepted feature greatly impacts relative false accept rate. We expect the false-accept rate to drop significantly once we collect a larger body of data and the voice channel is integrated with the gait subsystem.

Voice Analysis

Voice recognition was performed on speech data collected from 7 people. Each recording is about 1 minute long. We used 10 seconds of each recording for training purposes, and we tested our model on the remaining parts.

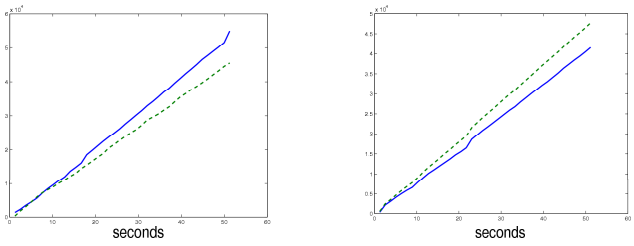

Fig. 3. Posterior probabilities of the target subject's test data (left) and impostor's test data (right) computed with target subject's model (solid line) and the impostor model (dashed line)

It can be seen from the graph (Fig. 3) that we are able to verify our target subject with about 10 seconds of speech data. As it was expected, the longer speech sample we are given the more reliable is the authentication. The speech was recorded in a relatively noise-free environment. Because of the low number of subjects and the controlled conditions, person identification using Gaussian Mixture Model was near perfect (over 98% accuracy). We used a mixture of five Gaussians in our subject model. Figure 4 shows the projections of those models into 3D space.

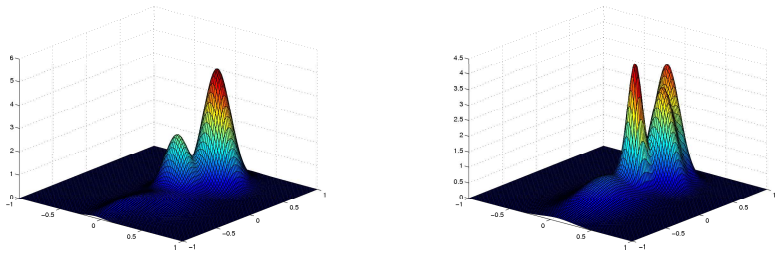

Fig. 4. Projections of twelve-dimensional mixtures of Gaussians into 3D space. The target model is on the left, and the impostor model is on the right

In general the accuracy results deteriorate with the number of people in the environment and the background noise. However, in an eldercare environment, the universe of individuals likely to come in contact with the subject is limited, typically consisting of family members and a small population of caregivers. Therefore, good accuracy results in our experiments are likely to be reproduced in real-world settings.

4 Summary and Future Work

We have built subject filtering module for our multimodal system for biometric analysis that relies on data from two audio and a vibration channel. We have shown that strong person verification can be performed using multimodal biometric system based on voice and gait. Our gait analysis subsystem indicates that the performance of the system improves by implementing information fusion of audio and vibration channels.

We introduced a novel approach in the gait recognition, namely the one based on audio and vibration of the foot impact with the floor. The preliminary results show that it is informative enough for our verification task.

The recognition accuracy of the voice channel was very high, though it could have been a result of a small data set and low-noise conditions during data acquisition. Thus we can cautiously expect high accuracy once we integrate it with the gait analysis subsystem.

Our models can be effectively used for the final stage of the research which is the passive monitoring of wellness. One of the approaches we are working on is to compare the current model state with the state which was archived a fixed time period ago. The model state is defined as a metric in the domain of parameters of both subsystems.

The preliminary results encourage further research in the multimodal biometric data fusion for recognition and monitoring. Current computational power at hand allows us to improve existing biometric recognition approaches by collecting biometric data through alternative channels, as we have done with audio- and vibration-based gait analysis. Finally, many of the advanced biometric recognition techniques which are currently used in security and authentication applications can also be used for health monitoring purposes.

References

1. Atal, B. (1976). "Automatic recognition of speakers from their voices," *Proceedings of IEEE*, **64**, 460-475.
2. Brunelli, R. and Falavigna, D. (1995). "Person identification using multiple cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**, no. 10, 955-966.
3. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society*, **39**, 1-38.
4. Domingos, P. and Pazzani, M. (1997) "Beyond independence: Conditions for the optimality of the simple bayesian classifier," *Machine Learning*, **29**, 103-130.
5. Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification.* Wiley, New York, 2 ed.
6. Gopinath, R. A. (1998). "Maximum likelihood modeling with Gaussian distributions for classification," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, **2**, 661-664.
7. Haykin, S. (1999). *Neural networks: A comprehensive foundation.* Prentice Hall, Upper Saddle River, 2 ed.
8. Hong, L., Jain, A. K., and Pankanti, S. (1999). "Can multibiometrics improve performance?," *Proceedings of AutoID'99: IEEE Workshop on Automated ID Technologies*, 59-64.
9. Jolliffe, I. T. (2002) *Principal Component Analysis.* Springer, New York, 2 ed.
10. Reynolds, D. A. and Rose, R. C. (1995). "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, **3**, 72-82.
11. Stevens, S. S. and Volkmann, J. (1940). "The relation of pitch of frequency: A revised scale," *American Journal of Psychology*, **53**, 329-353.