Multi-modal Person Recognition for Vehicular Applications

H. Erdoğan[1], A. Erçil[1], H.K. Ekenel[2], S.Y. Bilgin[1], İ. Eden[3], M. Kirişçi[1], and H. Abut[1,4]

[1] Sabancı University, Istanbul, 34956, Turkey
[2] University of Karlsruhe, Karlsruhe, 76131, Germany
[3] Brown University, Providence, RI, 02912, USA
[4] San Diego State University, San Diego, CA 92182, USA
{haerdogan, aytulercil, abut}@sabanciuniv.edu.tr

Abstract. In this paper, we present biometric person recognition experiments in a real-world car environment using speech, face, and driving signals. We have performed experiments on a subset of the in-car corpus collected at the Nagoya University, Japan. We have used Mel-frequency cepstral coefficients (MFCC) for speaker recognition. For face recognition, we have reduced the feature dimension of each face image through principal component analysis (PCA). As for modeling the driving behavior, we have employed features based on the pressure readings of acceleration and brake pedals and their time-derivatives. For each modality, we use a Gaussian mixture model (GMM) to model each person's biometric data for classification. GMM is the most appropriate tool for audio and driving signals. For face, even though a nearest-neighbor-classifier is the preferred choice, we have experimented with a single mixture GMM as well. We use background models for each modality and also normalize each modality score using an appropriate sigmoid function. At the end, all modality scores are combined using a weighted sum rule. The weights are optimized using held-out data. Depending on the ultimate application, we consider three different recognition scenarios: verification, closed-set identification, and open-set identification. We show that each modality has a positive effect on improving the recognition performance.

1 Introduction

Biometric person identification is a new and exciting research area which finds application in many different problems related to authentication, access control, keyless entry, and secure communications. Application of person and behavior identification in a vehicular environment has also attracted interest recently. This paper presents experiments for recognizing people in moving vehicles.

Due to competition in automotive industry, it is not too far when we will have cameras, microphones and various other sensors inside a vehicle that will gather and process multimedia data with the purposes of safer driving, improved comfort of driver and the passengers, and secure communications. Recognizing people in a car will be important to achieve the following benefits [1]:

N.C. Oza et al. (Eds.): MCS 2005, LNCS 3541, pp. 366 – 375, 2005.

1. Ensuring safety of the vehicle by requiring authorization before and/or during driving a car to make sure the current driver is an authorized driver,
2. Personalizing the vehicle suiting the driver's physical and behavioral characteristics, thereby, creating a comfortable, safe and efficient driving environment which minimizes distractions, and hence avoidance of many accidents attributed to driver distraction,
3. Providing safety to the vehicle, people, and goods in a commercial vehicle, via passive and active warning systems, even enabling authorities to disallow a driver who should not be or is not in a condition to be behind a wheel,
4. Opening opportunities to secure mobile transactions within a car, such as mobile banking, using biometric authentication.

There are serious challenges to person identification inside a car, especially if we are to assume no user cooperation. Over the past two decades, many algorithms, systems, and even technologies for speaker and face identification have been developed with varying degree of success (acceptable through excellent). Having been designed under idealized and controlled environments, however, both modalities suffer due to non-ideal conditions in real-world environments. In face recognition, for instance, change of illumination and pose, occlusions, facial expression, facial accessories, facial hair tend to deteriorate performance. For speaker recognition, external noise and channel effects, illnesses affecting the glottis and vocal tract, emotional speech may decrease performance. There are many studies to improve the performance of each modality within itself, such as to extract more robust features and to use more efficient normalization methods. Unfortunately, most of the methodologies under consideration are fairly mature and major breakthroughs are not forthcoming. Alternately, the research focus has shifted to the usage of multiple modalities together, so that when one of the modalities is not reliable or fails, other modalities can be relied upon.

In this paper, we attempt to use three different modalities, namely, speech, face and driving signals to recognize drivers of moving vehicles. We use MFCC features for speech, PCA features for face and the features extracted from the pressure readings of the acceleration and brake pedals and their derivatives. We combine information from each modality by computing a weighted sum of normalized modality scores. We determine the best weights by optimizing the verification performance on held-out[1] data. We consider three different types of person recognition: (*i*) verification, (*ii*) closed set identification, and (*iii*) open set identification, which will be explained in the next section.

We report our experimental results on a twenty people subset of the in-car corpus collected by the Center for Integrated Acoustic Information Research (CIAIR) [2]. We organize the paper in the following way. After introducing types of person recognition problems in section 2, we briefly introduce speaker and face recognition algorithms in sections 3 and 4. We explain how we used driving signals to recognize people in section 5. Next, we give details about our fusion algorithm. The experimental results are presented in section 7 and the conclusions are provided in section 8.

---

[1] The held-out data is a portion of available training data that is not used during training or testing, but used to adjust certain parameters of the recognition system. Sometimes held-out data is called validation data.

2 Problem Formulation

The task of recognizing people in vehicles is difficult for the following reasons:

- In vehicles, the subjects, especially the driver, are not expected to pose for the camera since their first priority is to operate the vehicle safely. Hence, there are large illumination and pose variations. In addition, partial occlusions and disguise are common.
- The quality of video is usually low, and due to the acquisition conditions and the physical constraints in positioning the camera, the face image sizes are smaller (sometimes much smaller) than the assumed sizes in most existing still image based face recognition systems.
- Speech acquisition in a car is prone to noise and channel distortions due to the engine and mechanical noise and reverberations in the vehicular chamber. For comfort and ease of use, far-talking microphones are employed instead of near-talking or head-set microphones, which decreases signal-to-noise ratio significantly and makes speaker recognition much more difficult.

Therefore, the use of multimodal biometrics becomes the most sensible route to follow for robust and reliable person identification inside a moving vehicle.

As in all other applications, the person recognition inside a car can be formulated as either a verification problem or an identification task. In the verification problem, a person's claimed identity is verified using her/his model in a known pool of subjects. On the other hand, one must be more careful in formulating an identification problem, which can be cast as either an open-set or a closed set identification problem. In the closed-set case, a reject scenario is not defined and an unknown subject is classified as one of the *N*-registered people. In the open-set case, the goal is to decide whether the person is among the registered people in the database or not. The system identifies the person if there is a match and otherwise rejects the claimed identity. Hence, the problem becomes an *N+1-class* identification problem, including a reject class. It is not difficult to see vehicle safety application can be addressed using an open-set identification scenario, while in-vehicle secure transactions application may be addressed under a verification task.

3 Speaker Recognition Mode

Speech signal is the most natural and non-invasive modality to identify a person in a vehicle. As in many other parametric speech processing applications, a set of features are extracted for each frame of speech over a short overlapping and advancing time window. It is worth noting that we preprocess speech signals to detect voice activity and extract features only from regions of audio where voice activity is present.

Features used for speaker recognition differ slightly from the ones used for speech recognition. In this study, we have used 12 coefficients of the Mel-frequency cepstral coefficients (MFCC) feature vector [3], i.e., in order to avoid dependence on acquired voice's energy, we have not included the energy coefficient. In this work, we did not use Δ and $\Delta\Delta$ features, which approximate first and second differences at the current frame respectively, as well, since their inclusion did not show noticeable improvement as reported in an earlier study [4].

MFCC features are obtained using a filterbank of overlapping triangular filters placed according to the critical bands of hearing [3]. The logarithms of filter output energies are computed. Then a DCT transform of these log-filterbank-energies is taken to de-correlate and reduce the dimension of the feature set as follows:

$$c_k = \sum_{j=1}^{N} m_j \cos\left(\frac{\pi k}{N}(j-0.5)\right), \tag{1}$$

where $\{c_k\}$ represent MFCC features and $\{m_j\}$ stand for log-filterbank-energies. These speaker features are considered as independent identically distributed random vectors drawn from a parametric probability density function (pdf). To model the pdf, Gaussian mixture models (GMM) are commonly used in speech processing community:

$$\mathrm{f}(\mathbf{x} \mid S_i) = \sum_{k=1}^{K} \gamma_k \ \mathrm{N}(\mathbf{x}, \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}). \tag{2}$$

Here $\mathbf{x}$ represents the feature vector, γ_k are mixture coefficients and $\mathrm{N}(\mathbf{x}, \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k})$ are individual Gaussians for representing a particular speaker S_i. For computational reasons, $\boldsymbol{\Sigma}_\mathbf{k}$ are chosen to be diagonal matrices. GMMs have been used in text-independent speaker recognition with great success [5]. A popular way of using GMMs in speaker recognition is to train a large background speaker model (say with 1024 Gaussians) and adapt this model to each speaker using that particular speaker's data. GMM training is performed via the EM algorithm [6].

In this paper, we train a GMM for each speaker from scratch and we use eight mixtures, which nevertheless gives satisfactory performance in this application. We had compared the performance of eight and sixteen mixtures in an earlier study [4] and obtained a better result using eight mixtures. During the testing phase, the per-frame log-likelihood value of observed data $(\mathbf{x}_j)_{j=1}^{N}$ under the model of a particular speaker S_i can be computed as:

$$L_i = \frac{1}{N}\sum_{j=1}^{N} \log \mathrm{f}(\mathbf{x}_j \left| S_i \right.) = \frac{1}{N}\sum_{j=1}^{N}\left(\log \sum_{k=1}^{K} \gamma_k \ \mathrm{N}(\mathbf{x}_j, \boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_k) \right). \tag{3}$$

We also train a background model, one more GMM, with twice the number of mixtures. Background GMM is required for normalization in likelihood ratio testing for speaker verification. The log-likelihood of the observed data under the background model, L_g can also be computed in a similar way. For verification task, the Bayesian decision amounts to the comparison of the log-likelihood-ratio, L_i-L_g to a threshold.

Robustness against noise can be an important issue in speaker recognition, especially if the training and testing conditions are mismatched. In our case, we have had the training and testing conditions matched. Hence, we did not perform any specific robustness algorithm such as feature and score normalization. In our future studies, we plan to include algorithms for robustness against noise and channel effects.

4 Face Recognition Mode

Among the plethora of face recognition methods, the paradigm based on face appearance data, template-based algorithms and their concomitant subspace versions, such as PCA and LDA methods are the most popular (see [7] for a comprehensive review). Since number of pixels in a face image can be rather large, it is reasonable to reduce feature dimension by projecting to a lower dimensional subspace. Thus, subspace projection techniques perform well for face recognition. Principal component analysis (PCA) is the most popular subspace projection technique used for face recognition [8-10].

PCA computes a linear transformation that maximizes the total scatter of the face images in the projected space. PCA aims to determine a new orthogonal basis vector set that best reconstructs the face images in the mean-squared error sense. These orthogonal basis vectors, also called eigenfaces, are the eigenvectors of the covariance matrix of the face images, associated with the highest eigenvalues.

In this study, we have trained a single Gaussian model for each person's face. Since we are using video signals, it is feasible to obtain many face images of a single person and it is feasible to use a statistical model for recognition. The decision making process is identical to the speech case after the statistical model is built.

5 Person Recognition Using Driving Signals

Can drivers be identified from their driving behavior? or equivalently, is the driving behavior a biometric trait? To answer this question, researchers at CIAIR and the authors of this paper have studied driving signals as measured by different sensors in the vehicle. Driving signals that were analyzed include pressure readings from accelerator and brake pedals, as well as the vehicle speed variations [11]. After trying Fourier analysis and multi-dimensional linear prediction techniques with limited success, both groups have employed GMM method to model driving signal characteristics. GMMs are successfully used for modeling speech signals in speaker recognition and are well-suited for application to driving signals as well. We believe this can be attributed to the fact that temporal characteristics of driving signals exhibit quasi-stationary behavior like speech. Smoothed and sub-sampled driving signals (acceleration and brake pedal pressures) and their first derivatives were used as features for modeling driving behavior of the drivers. Driving signals can be obtained by frequent sampling in time, thus we can collect ample data from a single person to train a statistical model. After feature extraction, the statistical modeling (driver/impostor models) part is just like the speech case. Similarly, we construct a GMM to model the driving features of each person and also train a background model.

6 Fusion

In this work, fusion of information from different modalities is performed at the matching score level, which is often called "decision fusion". We have used the weighted sum rule to combine scores from different modalities. As reported in literature [12, 13], the weighted sum rule is more robust against noise and other disturbances as compared to several other score combination rules, such as product rule, max rule and min rule, and often outperforms them.

An important aspect of classifier combination at the score level is to carefully normalize scores from each modality before the actual combination. Typical likelihood ranges for genuine and impostors could differ largely among modalities. Thus, log-likelihood-ratio scores from different modalities cannot be directly superimposed. Therefore, it is logical to normalize scores to make them compatible. One way to normalize scores is to use the mean and standard deviation of likelihood scores obtained from held-out validation data. Normalization can be performed using a sigmoid function which will map the scores to the (0,1) range.

$$S'_k = \frac{1}{1 + \exp\left(-(S_k - \mu)/\sigma\right)}. \tag{4}$$

Here S_k denotes the old log-likelihood-ratio score for the k^{th} modality, S_k' represents the new score. Furthermore, μ and σ are mean and standard deviation of old scores obtained on the validation set using all validation instances and all speaker models. In this work, we have used top $3N_t$ scores for N_t validation instances to compute the mean and standard deviation of scores, otherwise the mismatch scores (outnumbering genuine scores 19 to 1) would have dominated the statistics.

After normalization, we compute the weighted sum of new scores for each validation test case using the following formula:

$$S = \sum_{k=1}^{3} w_k S'_k. \tag{5}$$

We have chosen fixed weights w_k to minimize the verification equal error rate (EER) [2] on the validation data. The minimization is performed by exhaustively searching the weight space. After determining the optimal values for the weights on the validation data, we have employed them during testing phase for test data to compute overall final scores.

7 Experiments and Results

CIAIR at Nagoya University in Japan has been collecting an in-car speech database since 1999 with a data collection vehicle they have designed [2]. This vehicle supports synchronous recording of multi-channel audio data from 12 microphones that can be placed in flexible positions, multi-channel video data from 3 cameras and the vehicle related data such as the vehicle speed, the engine rpm, the steering wheel angle, acceleration and brake pedal pressures, where each channel is sampled at

[2] EER for verification is defined as the error rate when the false accept rate (FAR) is equal to the false reject rate (FRR) on the receiver operating characteristics curve which plots FRR versus FAR for different classification thresholds.

1.0 kHz. During the data collection stage, each subject has conversations with three types of dialogue systems. One is a human navigator, another is a Wizard of Oz system, and the last is a conversational system [2].

We have carried out person recognition experiments over a 20 person subset of the CIAIR database which consists of 812 drivers with well over a terabyte of data. We have used the camera facing the driver and the audio signal from the headset microphone for each person as video and audio sources, respectively. The faces were hand-cropped to 64x40 pixel size and non-silence audio sections were hand selected. We have smoothed and down-sampled the brake and acceleration pedal pressure readings by a factor of ten and their first derivatives to be the features for modeling the behavior of the driver. This resulted in four features at 100 Hz. Twelve static MFCC features (excluding c0) at 100 Hz were used as audio features. For faces, the PCA method was used to reduce the feature dimension to 20 for each image frame. The frame rate is 25 frames per second for the video.

From each driver, 50 image frames, 50 seconds of non-silence audio and around 600 seconds of driving signals were utilized. We extracted features from this dataset and divided all features into 20 equal length parts for each driver and modality and number the parts from one to 20. When we have formed the multimodal test-sets, we have assumed that each modality part was associated with the parts that have the same number in other modalities.

We have then performed a leave-one-out training procedure, where for each single testing part, seventeen parts were used for training and two parts were held-out for validation to optimize normalization parameters and fusion weights. This gave us 20 tests for each person (each time the training data is different although not independent), leading to 400 (20x20) genuine tests total. GMMs were used with eight, one and eight mixture components for speech, face and driving signals, respectively. Background GMM models were trained for each modality as well [6].

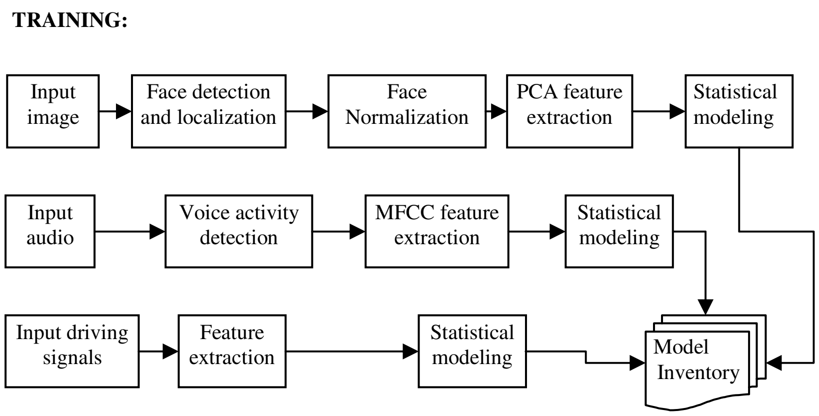

Fig. 1. System block diagram for training the multimodal driver recognition system

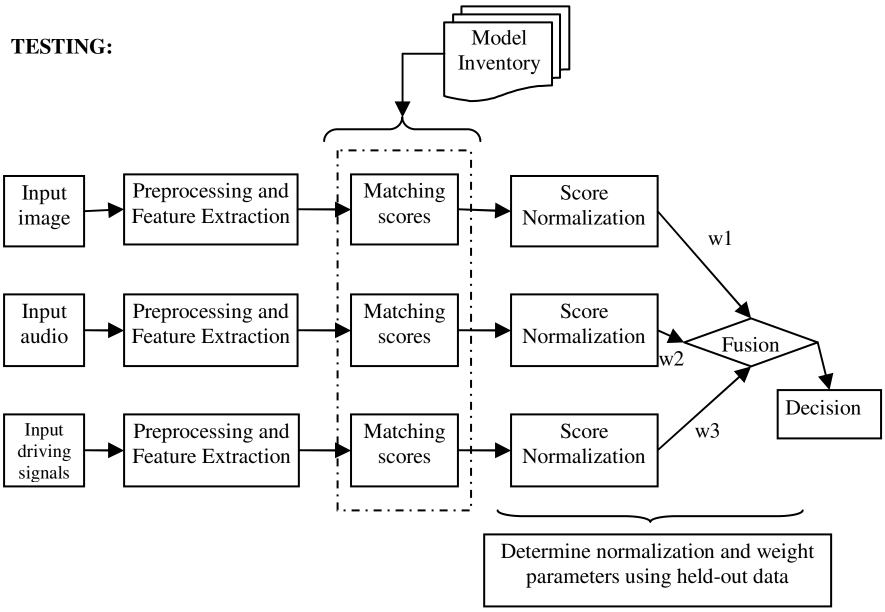

Fig. 2. System block diagram for testing the multimodal person recognition system

Block diagram of our training procedure is shown in Figure 1. Person recognition system is illustrated in Figure 2. We performed verification, closed set identification and open set identification tasks with the data. For verification, we have assumed each person's data as an impostor for the remaining 19 other drivers resulting at 7600 (19x20x20) impostor tests in total. For open-set identification tests, we leave one identity out at a time and perform open-set identification using the remaining 19 as genuine identities. We cycle through the set of identities to leave a different identity out each time to obtain 20 different testing setups for each test data. This procedure gives us 7600 genuine tests and 400 impostor tests. In this paper, we define EER for open-set scenario as the error rate when false-accept rate among the impostor attempts is equal to the sum of false-reject and false-classify rates among the genuine attempts.

The modalities were fused by the weighted score summation method mentioned earlier in section 6. Our findings from both the unimodal and multimodal performances are presented in Table 1.

The results from single-mode identification and verification are encouraging. As expected, audio-only yields the best performance since the speech samples were from the close-talking headset microphone. In a controlled lab environment face recognition algorithm has performed very successfully [1]. But in the CIAIR database, driver face segments were fairly small in comparison to other studies [4] and hence the results are expectedly not as good. We expect to get significantly higher results from face modality by using a custom-designed camera built-in to the visor which can be focused primarily on the face of the driver. The results based on analog driving signals are quite satisfactory and show improvement over an earlier study [11].

Table 1. Closed-set person identification, person verification and open-set person identification results

Modality	Weights	Closed-set ID (Accuracy %)	Verification (EER %)	Open-set ID (EER %)
A	Audio only	98.00	2.15	8.05
F	Face only	89.00	6.08	18.56
D	Driving only	88.25	4.00	21.06
A+D	(.62,.38)	99.25	0.83	3.75
F+D	(.43,.57)	98.00	1.62	8.86
A+F	(.63,.37)	99.75	0.50	1.75
A+F+D	(.47,.33,.20)	100.0	0.25	0.25

Pair-wise fusion scenarios result in significantly better performance over the face-only or driving-signals-only cases and even an incremental improvement is observed over the audio-only case. In many driver verification applications, an error rate of 0.5-1.62 percent would be satisfactory. For open-set identification, an EER rate of 1.75 percent, achieved by audio and face modalities, could be quite satisfactory as well.

As expected, the inclusion of all three modalities increases the performance of the person recognition system to an encouraging level. We believe that error rates of ¼ percent can bring most of the applications cited at the introduction section to reality and commercially viable systems can be built. Using multi-modal person recognition in a car is even more important than these results reveal, since any one of these modalities may fail or become impractical in certain cases, such as during driving at night or when there is presence of radio or other speakers in the vehicle.

However, we would like to point out that the results reported here are based on a relatively small data set and the investigators are experimenting with a much larger data set from the CIAIR corpus. We are also putting together a framework for a comprehensive and language/region-independent driver-specific data collection setup for the purposes of person recognition in a vehicle.

8 Conclusion

In this paper, we have introduced a multi-modal person recognition system that uses speech, face and driving signals for in-vehicle applications. It is interesting to note that, every modality has its own importance and improves the performance of the recognition system. Especially, it is interesting to see that driving signals are indicative of the person and those signals can be considered as a biometric trait which was not considered before.

We have obtained very encouraging results from a 20 person subset of the CIAIR database and have observed improvement in every multi-modal combination that we tried. These results show that, multimodal person recognition in a car is very promising. We conjecture that the improvement will be more important for adverse conditions when one of the modalities may become totally unreliable; nevertheless, it will still be possible to rely on the remaining modalities.

Acknowledgements

We would like to acknowledge Professor Kazuya Takeda of Nagoya University and his laboratory for providing the CIAIR database and Professor Fumitada Itakura of Meijo University, Nagoya, Japan for posing the original question and encouraging us to undertake the problem. Finally, we have been enjoying immensely our on-going collaboration with Professors A. Murat Tekalp, Engin Erzin, and Yücel Yemez of Koç University, Istanbul, Turkey.

References

[1] E. Erzin, Y. Yemez, A.M. Tekalp, A. Erçil , H. Erdogan, and H. Abut, “Multimodal Person Identification for Human Vehicle Interaction,” accepted for publication in the IEEE Signal Processing Magazine Special Issue on Man-Machine Communication, to appear September 2005.

[2] N. Kawaguchi, S. Matsubara, I. Kishida, Y. Irie, H. Murao, Y. Yamaguchi, K. Takeda and F. Itakura, “Construction and Analysis of the Multi-layered In-car Spoken Dialogue Corpus,” Chapter 1 in *DSP in Vehicular and Mobile Systems,* Springer, New York, NY, 2005.

[3] P. Mermelstein and S.B. Davis, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoustics, Speech and Signal Processing, vol. 28, pp. 357-366, August 1980.

[4] H.K. Ekenel, S.Y. Bilgin, I. Eden, M. Kirisçi, H. Erdogan and A. Erçil, “Multimodal Person Verification from Video Sequences,” Proceedings, SWIM 2004, Maui, HI, January 2004.

[5] D.A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” Speech Communications, 17, 91-108, 1995.

[6] A Dempster, N Laird, M Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm”, J. Royal Statistical Soc., 39, 1, 1978.

[7] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld "Face Recognition: A Literature Survey," ACM Computing Surveys, pp. 399-458, 2003.

[8] M. Turk and A. Pentland, “Eigenfaces for recognition,” Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 586.591, September 1991.

[9] Y.Y.J. Zhang and M. Lades, “Face recognition: Eigenface, elastic matching, and neural nets,” Proceedings of the IEEE, Vol. 85, no. 9, pp. 1423.1435, September 1997.

[10] W. Zhao, “Subspace Methods in Object/Face Recognition,” in Proc. Int. Joint Conf. on Neural Networks, 1999.

[11] K. Igarashi, C. Miyajima, K. Itou, K. Takeda, H. Abut and F. Itakura, “Biometric Identification Using Driving Behavior,” Proceedings IEEE ICME 2004, June 27-30, 2004, Taipei, Taiwan.

[12] J. Kittler, M. Hatef, R. Duin, and J. Matas, “On combining classifiers,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998.

[13] A. K. Jain, A. Ross, and S. Prabhakar, “An Introduction to Biometric Recognition,” IEEE Trans. On *Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics,* August 2003.