MULTIMODAL USER IDENTIFICATION FOR NETWORK-BASED INTELLIGENT ROBOTS

Keun-Chang Kwak
Department of Control, Instrumentation, and Robot Engineering, Chosun University
375 Seosuk-dong Dong-gu, 501-759 Gwangju, Korea

Keywords: Multimodal user identification, Face recognition, Speaker recognition, Human-robot intelligent, Network-based intelligent robots.

Abstract: This paper is concerned with multimodal user identification based face and speaker recognition for Human-Robot Interaction (HRI) under network-based intelligent robot environments. Face and speaker recognition are frequently used in conjunction with HRI that can naturally interact between human and robot. For this purpose, we present Tensor-based Multilinear Principal Component Analysis (TMPCA) and Mel-Frequency Cepstral Coefficients-Gaussian Mixture Model (MFCC-GMM) to recognize with face images and speech signals obtained through network transmission, respectively. Furthermore, we investigate network-based multimodal user identification for the near future study. The experimental results on face and speaker database with distant-varying reveal that the presented method shows good performance in network-based intelligent robot environments.

1 INTRODUCTION

Network-based intelligent home service robot what is called Ubiquitous Robotic Companion (URC), which exploits strong information technology infrastructure such as high-speed internet, is being used to develop the information service robots in Korea. URC is a new concept and network-based home service robot that will provide various advanced functions and services by adding the network to the existing robot, enhancing mobility and user interface considerably (Ha and Sohn, 2005). Recently, there has been a renewal of interest in Human-Robot Interaction (HRI) that can naturally interact between human and robot under network-based home service robot environments. The audiovisual HRI components include face detection, face recognition (Kim and Cha, 2007), face verification (Yun and Yoon, 2007), facial expression recognition, gesture recognition, action recognition, speaker recognition (Kwak and Yoon, 2007), sound source localization (Kwak and Kim, 2008) sound separation, and so on. Among various HRI components, we especially focus on multimodal user identification based on face and speaker recognition technique that is essential for network-based home service robots.

For this purpose, we develop a new approach to recognize simultaneously the successive frames with face images obtained through network transmission**.** Therefore, we elaborate on network-based face recognition with fast processing time based on Tensor-based Multilinear Principal Component Analysis (TMPCA) considering third-order tensor such as column and row of image as well as frame for human-robot interaction in network-based home service robot environments. On the other hand, we present text-independent speaker recognition based on Mel-Frequency Cepstral Coefficients-Gaussian Mixture Model (MFCC-GMM) with circular microphone array equipped with intelligent service robot. Furthermore, we investigate network-based multimodal user identification based face images and speaker's speech signals. This study is performed on the face and speaker database with distant-varying constructed in u-robot test bed environments to test and evaluate commercial robots.

2 FACE RECOGNITION

In this section, we focus on TMPCA for network-based face recognition. We perform feature extraction by determining a multilinear projection

Kwak K..
MULTIMODAL USER IDENTIFICATION FOR NETWORK-BASED INTELLIGENT ROBOTS.
DOI: 10.5220/0003574203370340
In *Proceedings of the 8th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2011), pages 337-340
ISBN: 978-989-8425-75-1

that captures most of the original tensor input variation. This is done by using the MPCA technique demonstrated in gait recognition (Lu and Plataniotis, 2008).

A N'th-order tensor is denoted as $A \in R^{I_1 \times I_3 \times ... \times I_N}$. It is addressed by N indices $i_n, n = 1,...,N$, and each i_n addresses the n-mode of A. The *n*-mode product of a tensor A by a matrix $U \in R^{J_n \times I_n}$, is denoted as $A \times_n U$.

From multilinear algebra, tensor A can be expressed as follows

$$A = S \times_1 U^{(1)} \times_2 U^{(2)} \times ... \times_N U^{(N)} \quad (1)$$

where $U^{(n)} = (u_1^n u_2^n ... u_{I_n}^n)$ is an orthogonal $I_n \times I_n$ matrix. Let $\{A_m, m = 1,...,M\}$ be a set of M tensor samples in $R^{I_1} \otimes R^{I_2} ... \otimes R^{I_N}$. The total scatter of these tensors is defined as

$$\Psi_A = \sum_{m=1}^{M} \left\| A_m - \bar{A} \right\|_F^2, \quad \bar{A} = \frac{1}{M} \sum_{m=1}^{M} A_m \quad (2)$$

Give all the other projection matrices $\tilde{U}^{(n)}$ consist of the P_n eigenvectors corresponding to the largest P_n eigenvalues of the matrix $\prod_{n=1}^{N} P_n$. A tensor can be projected to another tensor by N projection matrices as follows

$$Y = X \times_1 \tilde{U}^{(1)^T} \times_2 \tilde{U}^{(2)^T} \times ... \times_N \tilde{U}^{(N)^T} \quad (3)$$

For further details on the pseudo-code implementation of the MPCA algorithm, see (Lu and and Plataniotis, 2008).

3 SPEAKER RECOGNITION

In this section, the EPD (Endpoint Detection) algorithm is performed to analyze speech signal obtained from speaker. Here the speech signal is detected by log energy and zero crossing. After detecting signal, the feature extraction step is performed by six stages to obtain MFCC. These stages consist of pre-emphasis, frame blocking, hamming window, FFT (Fast Fourier Transform), triangular bandpass filter, and cosine transform. For simplicity, we use 11 MFCC parameters except for the first order. In what follows, we construct GMM (Reynolds and Rose, 1995) frequently used in conjunction with text-independent speaker recognition to represent speaker's individual model in robot environments. A Gaussian mixture density is represented by a weighted sum of M component densities. The complete Gaussian mixture density is parameterized by mixture weights, mean vectors, and covariance matrices from all component densities. For speaker identification, each speaker is represented by a GMM and is referred to as model λ. A group of *S* speakers, $S = \{1,2,...,K\}$ is presented by GMM's $\lambda_1, \lambda_2, \cdots, \lambda_K$. The objective is to find the speaker model which has the maximum a posteriori probability for a given observation sequence $(x_1, x_2, \cdots, x_T)$.

$$\hat{S} = \arg\max \sum \log(x_t \mid \lambda_k) \quad (4)$$

4 NETWORK-BASED MULTIMODAL USER IDENTIFICATION

Transmission rate is an important issue to perform human-robot interaction under network-based intelligent service robots. Because a low-price robot has the limitation of computation, our system usually transmit image and speech signals to the server, in which application programs such as speech recognition, speaker recognition, and face recognition are executed and their results are transmitted to a robot client. MIM board developed by ETRI was designed to be able to transmit speech signals and images simultaneously. The type of image is Jpeg (320*240) and speech signal is sampled by 16K rate with 16bit size. According to TCP transmission protocol, MIM board can transmit 6~8 frames per second (492~655Kbps) for an image and 256Kbps for speech, respectively. Generally, TCP/IP protocol can transmit about 10M~12M bps (Kim and Yoon, 2007).

Figure 1 visualizes scenario network-based multimodal user identification system based on face and speaker recognition through face images and speech signals transmitted from robot. We shall complete a fusion scheme for multimodal user identification in the near future.

5 EXPERIMENTAL RESULTS

This section reports on the experiments and draws conclusions as to the performance of the presented approach. Face and speaker database used in this

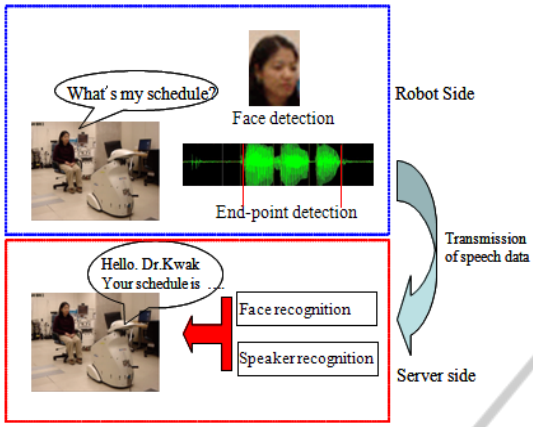


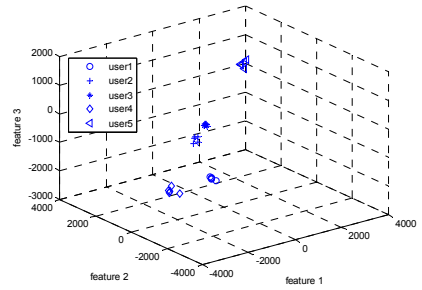
Figure 1: Scenario of network-based multimodal user identification system.

study were constructed in u-robot test bed as like home environments. We used frontal face video sequences with distant-varying in 1 m and 2 m, respectively. The face database includes 300 different images coming from 10 individuals. Among total database, we used 50 face images obtained from 1 meter for enrolment and the remaining 250 face images from 1 meter and 2 meter are used for test stage, respectively. We suppose that these frames were transmitted by network under network-based robot environments. Here training set includes 50 groups with 10 successive frames resized by 45x40. The test set consists of 50 groups and 200 groups with 10 successive frames for 1 m and 2 m, respectively. Each image was digitized and resized by a 45x40 whose gray levels ranged between 0 and 255. Figure 2 shows some example of original and detected face images obtained in distant-varying environments.

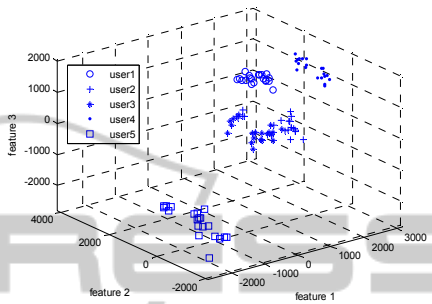
(a) 1m (b) 2m

Figure 2: Face images and detected results.

Figure 3 visualizes the distribution of feature vector of five robot's users for 1 m and 2 m in test set, respectively. Here, we only considered three feature points. As shown in Figure 3, we found that the presented method has a good class discriminability of the feature points in 1 meter. However, the feature points in 2 meter are a little overlapped for some classes due to distant-varying. The experimental results showed good performance (1m: 100%, 2m: 80%) in network-based intelligent robot environments



(a) 1m



(b) 2m

Figure 3: Distribution of feature vectors for test set.

On the other hand, we used speaker database to evaluate the identification performance of the presented speaker recognition system. Figure 4 shows robot platform with eight microphones and sound board, which is URC-based intelligent service robot developed by ETRI. Here sound board with eight channels. The database is constructed by audio recording of 10 speakers. The data set consists of 30 sentences for each speaker and channel. We divided the speech data into training (20 sentences × 10 people × 2 distances (1,2m) × 8 channels) and test data sets (10 sentences × 10 people × 2 distances (1,2m) × 8 channels). For simplicity, we use speaker database consisting sentences for only single microphone. Figure 5 shows same speech signals recorded from microphones equipped in robot platform. The recording was done in an office environment. We performed text-independent speaker identification under robot environments. The experimental results showed a good identification performance (1m: 94.5%, 2m: 94.8%).

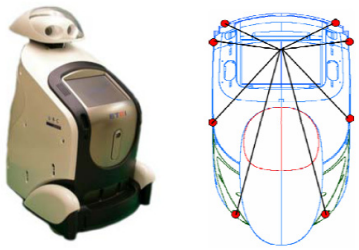
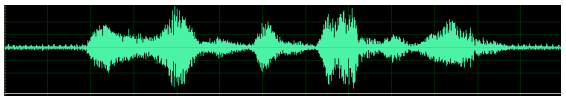
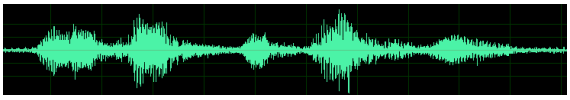


Figure 4: Robot platform and arrangement of microphone.



(a) 1m



(b) 2m

Figure 5: Speech signals obtained in 1m and 2m.

6 CONCLUSIONS

We have developed the face and speaker recognition system for multimodal user identification with the aid of TMPCA and MFCC-GMM under network-based home service robot environments, respectively. Furthermore, we have used the low-price camera and microphones for the commercialization of network-based home service robots. The experiments were performed on the face and speaker database constructed in u-robot test bed as like home environments. The presented method could effectively recognize in network-based environment, which is useful not only for intelligent home robots but also for biometrics and digital home networks. We shall study an efficient fusion scheme for network-based multimodal user identification in the near future.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology. (20110003296)

REFERENCES

Reynolds, D. A., Rose, R. C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83.

Kwak, K. C., Kim, H. J., Bae, K. S., Yoon, H. S., 2007. Speaker identification and verification for intelligent service robots. In *International Conference on Artificial Intelligence (ICAI2007)*, Las Vegas, May.

Ha, Y. G., Sohn, J. C., Cho, Y. J., and Yoon, H., 2005. Towards ubiquitous robotic companion: Design and implementation of ubiquitous robotic service framework. *ETRI Journal*, vol. 27, no. 6, pp. 666-676.

Kim, D. H., Lee, J., Yoon, H. S., and Cha, E. Y., 2007. A non-cooperative user authentication system in robot environments. *IEEE Consumer Electronics*, vol. 53, no. 2, pp. 804-811.

Yun, W. H., Kim, D. H., and Yoon, H. S., 2007. Fast Group verification system for intelligent robot service. *IEEE Trans. on Consumer Electronics*, vol. 53, no. 4, pp. 1731-1735.

Ji, M., Kim, S., and Kim, H., 2008. Text-independent speaker identification using soft channel selection in home robot environments. *IEEE Trans. on Consumer Electronics*, vol. 54, no. 1, pp. 140-144, 2008.

Lu, H., Plataniotis, K. N., and Venetsanopoulos, A. N., 2008. MPCA: Multilinear principal component analysis of tensor objects. *IEEE Trans. on Neural Networks,* vol. 19, no. 1, pp. 18-39.

Kwak, K. C., and Kim, S. S., 2008. Sound source localization with aid of excitation source information in home robot environments. *IEEE Trans. on Consumer Electronics*, vol. 54, no. 2, pp. 852-856.

Kim, H. J., Lee, J. Y., Kwak, K. C., and Yoon, H. S., 2007. Network-based voice component framework for human-robot interaction. *International Symposium on Communications and Information Technologies (ISCIT 2007)*, pp. 1546-1550.