Engineering

Industrial & Management Engineering fields

Okayama University *Year* 2006

A Real-life Test of Face Recognition System for Dialogue Interface Robot in Ubiquitous Environments

Fumihiko Sakaue* Makoto Kobayashi† Tsuyoshi Migita‡

Takeshi Shakunaga** Junji Satake††

*Nagoya Institute of Technology

†Okayama University

‡Okayama University

**Okayama University

††National Institute of Information and Communications Technology

A Real-life Test of Face Recognition System for Dialogue Interface Robot in Ubiquitous Environments

Fumihiko SAKAUE* Makoto KOBAYASHI Tsuyoshi MIGITA Takeshi SHAKUNAGA
Okayama University
3-1-1, Tsuhima-naka, Okayama-shi, 700-8530

Junji SATAKE
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289

Abstract

This paper discusses a face recognition system for a dialogue interface robot that really works in ubiquitous environments and reports an experimental result of real-life test in a ubiquitous environment. While a central module of the face recognition system is composed of the decomposed eigenface method, the system also includes a special face detection module and the face registration module. Since face recognition should work on images captured by a camera equipped on the interface robot, all the methods are tuned for the interface robot. The face detection and recognition modules accomplish robust face detection and recognition when one of the registered users is talking to the robot. Some interesting results are reported with careful analysis of a sufficient real-life experiment.

1. Introduction

Researches on unibiquitous computing have been discussed actively to enable a user to obtain high-level services from networked appliances without any explicit manipulation of the computers. In the present paper, we focus on a face recognition system which is utilized for an dialogue interface robot in a ubiquitous environment that were built in a usual apartment. The ubiquitous environment has various sensors and digital appliances networked together, and the home itself can understand inputs from the sensors as well as the residents' behaviors, situations and their social relations. In order to facilitate such intelligent environments, we have implemented a ubiquitous environment, in which the whole system is considered to be a mother and the dialogue interface robot is regarded as a child. Combining the interface robots and the other system components, the system can provide high-quality services to the user. Thus, the dialogue interface robot has to recognize requests from a user based on the audio input, and the robot has to answer the user's query using its voice systhesis capability if required. In addition, the robot has to identify the speaker based on the visual input in order to utilize the knowledge of the user's behavior and/or the situation.

*Fumihiko SAKAUE is currently with Nagoya Institute of Technology.

2. Face Recognition System in a Dialogue Interface Robot

We have implemented a face recognition system to be used in a dialogue interface robot [6] shown in Fig. 1. The height of this robot is 25cm, thus the robot can be put on various places. The robot is equipped with a USB camera (Kanebo KBCR-M01VU-RUB03) and a microphone (Sennheister ME105) for visual and audio interfaces with users. The camera captures color images of 240×320 pixels, and which are used to recognize the person who is having a conversation with the robot. This camera is used with a wide conversion lens (KenkoDIGITAL MPL-WA) to expand the field of view. Since this lens causes the images to be distorted, the images should be rectified [7] before processed by the face recognition system.

Several robots of this capability are installed in the *ubiquitous home*[6] environment, which is an apartment-like suite of rooms built in a research institute. A family can actually live in the suite. In addition, in the rooms, there are a number of sensors and digital appliances as well as other kinds of facilities which are networked together, and these are used for technical researches of ubiquitous environments. The interface robots serve as an interface between users and the computers. For example, a robot recognizes a user, and recommends his/her favorite TV program.

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

Figure 1. Dialogue interface robot.

Living room Kitchen

Figure 2. Two robots in living room and kitchen.

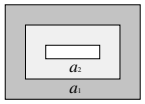

Figure 3. Square Separability Filter.

Figure 4. Examples of face detection result.

Figure 5. Example of eigenface: Mean (left) and principal vectors.

Figure 2 shows example scenes where a user has a conversation with the robot in the living room and in the kitchen.

We assume a distance between a user and the robot is from 30 to 80 cm, which is the distance the robot can hear the user's usual voice as well as his/her whispering. We also assume a user talks to the robot from the position just in front of the robot, thus we only have to consider a frontal face recognition. The robot should recognize 3 through 7 persons that seem the typical number of persons in a family.

3. Face Recognition Method

3.1. Face Detection by Square Separability filter and Eigenface

3.1.1. Eye detection by Square Separability Filter

In this section, a face recognition method is shown for the face recognition system. At first, the face detection method consists of two processes. Let us describe details of the processes.

In the first process of the face detection, possible eye positions are detected utilizing a separability filter[1]. A form of the filter is illustrated in Fig. 3. A separability S between area a_1 and area a_2 is given by

$$S = \frac{n_1(\overline{p_1} - \overline{p})^2 + n_2(\overline{p_2} - \overline{p})^2}{\sum_{i=1}^{N} p_i^2 - (n_1 + n_2)\overline{p}^2}, \quad (1)$$

where n_1 and n_2 are the numbers of pixels in exclusive areas a_1 and a_2, p_i is a pixel value of the i-th pixel in $a_1 \cup a_2$ and $\overline{p_1}, \overline{p_2}, \overline{p}$ are the mean pixel values over areas a_1, a_2 and $a_1 \cup a_2$, respectively. The filter is applied for an input image while changing the size in order to adapt the image scale. If the separability is larger than a threshold, the position is selected as a candidate of eye position. The separability filter can be efficiently implemented using an integral image.

3.1.2. Face Detection using Eigenface

In the second process, candidate face positions are estimated from possible combinations of eye candidates. In the estimation, some candidates are rejected due to invalid combinations. For example, a face candidate is rejected when the two eyes lie in a vertical line. Only the valid face candidates are evaluated in the image domain. Based on the positions of eye candidates, an input image is converted to 32 by 32 pixels so that all of the image has eyes in the same coordinates as shown in Fig. 4. In the conversion, all the pixel values are normalized so that the summation of them is 1. Let $\mathbf{x}$ denote a converted image.

Each converted image, $\mathbf{x}$, is then compared to an eigenface in order to judge whether the image is a face, or not. The eigenface is constructed by principal component analysis (PCA) from an image set which consists of a lot of facial images taken under various lighting conditions. In the current implementation, the image set comprises 50 persons of 24 lighting conditions and does not include any examinees of the real-life test. Let $\langle \overline{\mathbf{x}}, \Phi_m \rangle$ denote the eigenspace which has a mean vector $\overline{\mathbf{x}}$ and a matrix Φ_m of which each column represents an eigenvector. An example of the eigenface is shown in Fig. 5.

Each converted image is compared to the eigenface by parallel partial projections (PPP)[2]. The PPP can work more robustly under complex lighting conditions than a simple projection or conventional robust projections. The scheme of parallel partial projections is illustrated in Fig. 6. In the scheme, a converted image is divided into square regions and each region is partially projected onto the eigenface by

$$\tilde{\mathbf{x}}_i^* = (P_i \tilde{\Phi}_m)^+ \mathbf{x}, \quad (2)$$

where P_i is a 1024x1024 diagonal matrix that specifies the i-th region. In P_i, each diagonal term is 1 or 0, which indicates whether the pixel is effective (1) or ineffective (0) for

0-7695-2521-0/06 $20.00

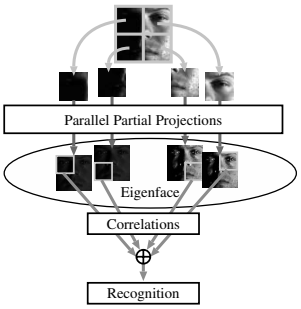

Figure 6. Face Evaluation by Parallel Partial Projections.

the i-th partial projection. In Eq. (2), $\tilde{\Phi}_m = [\Phi_m \ \overline{\mathbf{x}}]$, and $A^+ = (A^T A)^{-1} A^T$.

The converted image is, then, evaluated by

$$v = \sum_{i=1}^{M} C(P_i \mathbf{x}, P_i \tilde{\Phi}_m \tilde{\mathbf{x}}_i^*), \tag{3}$$

where M is the number of regions and $C(\mathbf{a}, \mathbf{b})$ is calculated by

$$C(\mathbf{a}, \mathbf{b}) = \frac{\sum_{j=1}^{N} (a_j - \overline{a})(b_j - \overline{b})}{\sqrt{\sum_{j=1}^{N} (a_j - \overline{a})^2 \sum_{j=1}^{N} (b_j - \overline{b})^2}}, \tag{4}$$

when a_j and b_j are j-th components of $\mathbf{a}$ and $\mathbf{b}$, respectively, and N is the number of pixels in the regions. In the equation, $\overline{a} = (1/N) \sum_{j=1}^{N} a_j$ and $\overline{b} = (1/N) \sum_{j=1}^{N} b_j$.

A converted image is selected as a face candidate when its v-value is larger than a threshold and the image is recognized by a face recognition method as described in **3.2**.

3.2. Face Recognition by Decomposed Eigenface Method

The detected facial images are recognized by the decomposed eigenface method[4]. That is, input images are decomposed into independent two components, and the two components are evaluated independently by a conventional eigenspace method. It is considered that a conventional eigenface is decomposed into the two independent eigenspaces as shown in Fig. 7. The eigenface decomposition can be accomplished by a canonical space (CS) [4] which is a low dimensional eigenspace. However, the eigenface decomposition also can be done by Gaussian filter[3].

3.2.1. Eigenface Decomposition by Canonical Space

Eigenface decomposition by the canonical space is shown in this section. Here, the canonical space (CS) is an eigenspace

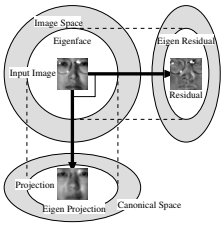

Figure 7. Decomposition of Eigenface

Figure 8. Examples of image decomposition: input images (upper row), projections (middle row) and residuals (lower row).

which is constructed from a lot of facial images taken under various lighting conditions. Let $\langle \overline{\mathbf{x}}_{cs}, \Phi_{cs} \rangle$ denote the canonical space which has a mean vector $\overline{\mathbf{x}}_{cs}$ and principal vectors Φ_{cs}. The canonical space decomposes an input image $\mathbf{x}$ into a projection $\mathbf{x}^{\$}$ and a residual $\mathbf{x}^{\sharp}$ by

$$\mathbf{x}^{\$} = \Phi_{cs} \Phi_{cs}^T (\mathbf{x} - \overline{\mathbf{x}}_{cs}) + \overline{\mathbf{x}}_{cs} \tag{5}$$

and

$$\mathbf{x}^{\sharp} = \mathbf{x} - \mathbf{x}^{\$}. \tag{6}$$

Examples of the projection and the residual are shown in Fig. 8.

The projection $\mathbf{x}^{\$}$ includes canonical information because it is represented in the CS. Since the CS can cover changes in geometric and in photometric aspects of faces, the canonical information includes those information. On the other hand, the residual includes salient individuality and noise because it is not represented in the CS.

(1) Registration stage A registration stage of the decomposed eigenface method is specified as follows: Let $\{\mathbf{x}_{pi}\}$ denote the registered images for a person p. The images are decomposed into $\{\mathbf{x}_{pi}^{\$}\}$ and $\{\mathbf{x}_{pi}^{\sharp}\}$ by Eqs. (5) and (6). Here, an eigenspace, called the eigen projection and denoted by $\langle \overline{\mathbf{x}}_p^{\$}, \Phi_p^{\$} \rangle$, is constructed by PCA from projections. At the same time, the other eigenspace, $\langle \overline{\mathbf{x}}_p^{\sharp}, \Phi_p^{\sharp} \rangle$, which is called eigen residual, is also constructed by PCA from residuals. The both eigenspaces are constructed for all registered persons.

(2) Recognition stage A face recognition algorithm is constructed in the conventional way on these two sets of eigenspaces. At first, an input image is decomposed into a projection $\mathbf{x}^{\$}$ and a residual $\mathbf{x}^{\sharp}$. They are, then, compared

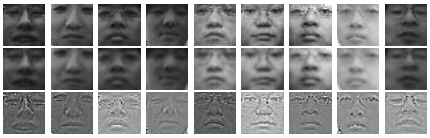

Figure 9. Examples of image decomposition by Gaussian filter: input images (upper row), Gaussian components (middle row) and residuals (lower row).

with eigenspaces of each person p independently by

$$C_p^{\$} = C(\mathbf{x}^{\$}, \Phi_p^{\$}\Phi_p^{\$^T}(\mathbf{x}^{\$} - \overline{\mathbf{x}}_p^{\$}) + \overline{\mathbf{x}}_p^{\$}) \tag{7}$$

and

$$C_p^{\sharp} = C(\mathbf{x}^{\sharp}, \Phi_p^{\sharp}\Phi_p^{\sharp^T}(\mathbf{x}^{\sharp} - \overline{\mathbf{x}}_p^{\sharp}) + \overline{\mathbf{x}}_p^{\sharp}). \tag{8}$$

A summation of the two correlations, $C_p^{\$} + C_p^{\sharp}$, indicates an evaluation value for p. Finally, the input image is recognized as a person which provides the highest evaluation when it is higher than a threshold. Otherwise, the input image is rejected since the person is not similar to any registered persons.

3.2.2. Eigenface Decomposition by Gaussian Filter

Eigenface decomposition could be done through Gaussian filtering instead of using the CS. This approach is similar to that of self-quotient image (SQI) [5] since both the methods utilize a Gaussian filter. However, the SQI extracts only the component that is insensitive to illumination, while our method decomposes an image into two components.

Let G formally denote an $N \times N$ matrix that works as the Gaussian filter. Then, the decomposition of image $\mathbf{x}$ into a Gaussian image $\mathbf{x}_G^{\$}$ and its residual $\mathbf{x}_G^{\sharp}$ by the Gaussian filter can be formulated as

$$\mathbf{x}_G^{\$} = G\mathbf{x} \tag{9}$$

and

$$\mathbf{x}_G^{\sharp} = \mathbf{x} - \mathbf{x}_G^{\$}. \tag{10}$$

Two eigenspaces are also constructed from the projections and residuals. They are constructed and evaluated by a similar way as described in **3.2.1**. Examples of the image decomposition by Gaussian filter are shown in Fig. 9.

The decomposed eigenface method can be combined with the parallel partial projections[2]. It is reported that the combined method works better than the original method[3].

4. Real-life Experiments

4.1. Experimental environments

A real-life test of the proposed face recognition system was performed in the ubiquitous home environment[6]. A family consisting of a husband, his wife and their 14-year-old daughter, spent 15 days in the experimental site. In the experimental site, they lived their ordinary lives as if they lived in their own home. Five sets of the interface robots were set in the living room, kitchen, bed room, study room, and entrance hall, respectively, and they were mainly used for conversation with a person near the interface robots. Each interface robot is equipped with a camera that is connected to a PC through a network, and the face recognition system works on the PC. For the face recognition, a set of face images were registered for each person in the face registration stage. Size of registered images is normalized to 32×32.

In the experiment, the face recognition system autonomously kept detecting a face that appeared in front of the interface robot, and discriminating the face. Since the family lived their ordinary lives in the experimental site, and the camera had only a narrow field of view, the face recognition system could not detect anybody when a person was out of the view field in their ordinary lives. Thus, almost all images included no face at all, and captured nothing valuable. In this situation, we should thoughtfully consider how to effectively analyze the experimental results.

Only when a person happened to appear in front of the interface robot and he/she would like to speak to the robot, the detection system could effectively detect his/her face. In the case, each person was trying to speak to the interface robot and the face recognition system worked for the person identification. Since both the audio and video data had been recorded with time stamp, we can detect each speech session and analyze what was detected while a person was speaking to the robot. This means that we can effectively evaluate the performance of face recognition over all the speech sessions. We have decided the face recognition results are evaluated only for dialogue sessions in the performance evaluation. The evaluation scheme is consistent with the objective of the interface robot since the robot is programmed to make a friendly Hello with using his/her name when the face recognition system can identify the person in front of the robot. Evaluation of the face recognition is accomplished by comparison of the video images and the recognition result in each dialogue session. Each session is defined by a set of the beginning and the end of each dialogue, and a session includes 5 through 20 images since a dialogue session is 1 through 3 seconds and the camera can take 5 through 7 images per second. We have evaluated our experimental results in these specifications.

Table 1. Comparison among the robot locations [%]: (a) misidentification rate and (b) rejection rate.

		CS		Gaussian	
location	#scenes	(a)	(b)	(a)	(b)
living	1115	3.6	42.6	4.5	10.6
kitchen	567	4.4	16.5	0.0	4.8
entrance hall	70	0.0	81.4	3.3	57.1
bed room	75	4.8	45.3	3.9	32.0
study room	12	0.0	50.0	0.0	41.7

4.2. Comparison of CS-decomposition and Gaussian-decomposition

We have compared the recognition rates between the CS-decomposition and Gaussian-decomposition. In this experiment, three persons and an additional person were registered a priori. The additional person to the family is an assistant person who sometimes comes into the experimental site. However, since he is not an examined, he had never appeared in any dialogue session. For each person, face images were taken in the living room and in the kitchen, respectively, for the registration.

Changing the rejection threshold, we have calculated the misidentification rates and the rejection rates for both the CS-decomposition and the Gaussian-decomposition. Here, the rejection rate means a rate of misdetection when a person is in front of the robot. Figures 10 and 11 show relations of the rejection rates and misidentification rates for the CS-decomposition and the Gaussian-decomposition, respectively. In these figures, the system performance is better when the characteristic curve approaches to the origin. Thus, the Gaussian-decomposition is better than the CS-decomposition.

Table 1 shows detailed results of the experiment for each interface robot. The left column shows the locations of each robot, and the second column shows how many dialogue sessions were taken during the experiment. The next two columns show both the misidentification rate and the rejection rate when the CS-decomposition is used for the decomposition, where the rejection rate shows a minimum one when the misidentification rate stays under 5 %. Since the misidentification rates are almost equal, the lower rejection rate means the better system performance in the table.

Table 1 indicates that the Gaussian-decomposition stably provided good results in the living room. Figures 10 and 11 also show that the Gaussian-decomposition has better characteristics than the CS-decomposition. These results mean that the canonical space cannot sufficiently cover a wide variety of lighting environments in the experimental site. While the canonical space is constructed from a lot

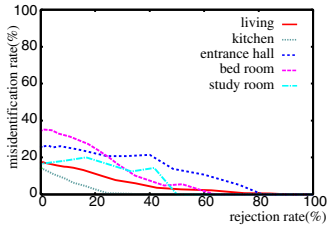

Figure 10. Relation of misidentification rate and rejection rate (CS-decomposition)

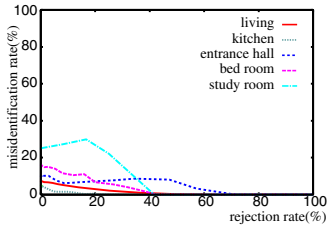

Figure 11. Relation of misidentification rate and rejection rate (Gaussian-decomposition)

of face images, they were taken with moving a single light source in a dark room. On the other hand, real lighting environments often include multiple light sources. The critical difference often affects face registration. Some important facial features were not properly decomposed by the projection, and this resulted in a bad eigenspace construction and consequently in a bad recognition rate. On the other hand, the Gaussian decomposition can learn face images as they were in the real site since no projection is involved in the registration stage. This natural coding facilitates better recognition performance than the CS-decomposition.

4.3. Comparison among different registrations

An additional experiment was accomplished for all the dialogue sessions in the living room and in the kitchen. In the experiment, registered images were selected in three ways: (1) Only the images taken in the living room were used for the registration. (2) Only the images taken in the kitchen were used for the registration. (3) Both of them were used for the registration. Figures 12-15 show the results of the misidentification rates and the rejection rates for the CS-decomposition and the Gaussian decomposition.

The living room and the kitchen have very different lighting conditions. Faces were illuminated by a front light in the living room, while faces were illuminated by a side light in the kitchen. The considerably different lighting conditions result in the bad results in the living room. The results

Figure 12. Comparison among different registrations(living room, CS-decomposition)

Figure 13. Comparison among different registrations(living room, Gaussian-decomposition)

Figure 14. Comparison among different registrations(kitchen, CS-decomposition)

Figure 15. Comparison among different registrations(kitchen, Gaussian-decomposition)

for the living room show that sufficient identification cannot be accomplished when only the kitchen images were registered. On the other hand, the results for the kitchen were not so bad even when only the living-room images were registered.

When images taken in the living room and in the kitchen were used for the registration, both the misidentification rate and the rejection rate decrease in comparison with the other two cases. This suggests that the system performance can be much more improved if faces can be registered in real and many different lighting conditions. However, increase of the registered images may result in lack of usefulness of the system. The efficiency and the usefulness should be balanced in real applications.

5. Conclusion

This paper discussed a face recognition system for a dialogue interface robot that really works in ubiquitous environments and reported an experimental result of real-life test in a ubiquitous environment. Some interesting results have been reported with careful analysis of a sufficient real-life experiment.

Acknowledgment

This work has been supported in part by a Grant-In-Aid for Scientific Research (No.15300062) from the Ministry of Education, Science, Sports, and Culture of Japan.

References

[1] K. Fukui, "Edge extraction method based on separability of image features," *IEICE Trans. Inf. & Sys.*, vol. E78-D, no. 12, pp. 1533–1538, 1995.

[2] F. Sakaue and T. Shakunaga, "Face recognition by parallel partial projections," *Proc. ACCV2004*, vol. 1, pp. 144–150, 2004.

[3] F. Sakaue and T. Shakunaga, "Combination of projectional and locational decompositions for robust face recognition," *Proc. IEEE Intl. Workshop on Analysis and Modeling of Faces and Gestures*, pp. 407–421, 2005.

[4] T. Shakunaga and K. Shigenari, "Decomposed eigenface for face recognition under various lighting conditions," *Proc. CVPR2001*, vol. 1, pp. 864–871, 2001.

[5] H. Wang, S. Li, and Y. Wang, "Face recognition under varying lighting conditions using self quotient image," *Proc. FG2004*, pp. 819–824, 2004.

[6] T. Yamazaki, "Human action detection and context-aware service implementation in a real-life living space test bed," *Proc. 2nd Intl. IEEE/Create-Net Conf. Testbeds and Research Infrastructures for the Development of Networks and Communities (TridentCom 2006)*, 2006.

[7] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. PAMI*, vol. 22, no. 11, pp. 1330–1334, 2000.