Tiny and Blurred Face Alignment for Long Distance Face Recognition

Kyu-Dae Ban, Jaeyeon Lee, DoHyung Kim, Jaehong Kim, and Yun Koo Chung

Applying face alignment after face detection exerts a heavy influence on face recognition. Many researchers have recently investigated face alignment using databases collected from images taken at close distances and with low magnification. However, in the cases of home-service robots, captured images generally are of low resolution and low quality. Therefore, previous face alignment research, such as eye detection, is not appropriate for robot environments. The main purpose of this paper is to provide a new and effective approach in the alignment of small and blurred faces. We propose a face alignment method using the confidence value of Real-AdaBoost with a modified census transform feature. We also evaluate the face recognition system to compare the proposed face alignment module with those of other systems. Experimental results show that the proposed method has a high recognition rate, higher than face alignment methods using a manually-marked eye position.

Keywords: Face alignment, face detection, face recognition, human robot interaction.

I. Introduction

Recently, there has been growing interest in human-robot interaction (HRI). Target-person identification is the first crossroad of HRI. Many researchers have recently investigated recognition methods for person identification, such as face, speaker, and gait recognition. Among these recognition methods, face recognition is the most frequently explored modality [1]. Most of the current face recognition methods are based on facial appearance [2]. In these face recognition methods, the face component positions of the probed face image have to be exactly the same as the gallery face image. The detected face, therefore, must be aligned before the recognition. However, it is difficult to align a face that has been captured from a robot camera.

A robot platform design is restricted due to insufficient inner space and production costs, so home-service robot manufacturers may prefer to use cheaper cameras, such as webcams. An inexpensive camera may have a small charge-couple device (CCD) or CMOS sensor and a lens with a small aperture. Thus, it is difficult to expect good quality in a face image taken by a robot. The sizes of faces in captured images are often small because of the distance between the robot and its user is always changing and can be long. In addition to cheap cameras, a low-resolution image, such as 320×240, is preferred because of limited computing power on the robot. These constraints also occurred in the server client model. We want a complex process, such as face and speaker recognition, to be performed at a server facility connected with a wireless network. Transmitting high-resolution image data to a robot's server can cause communication traffic.

The most traditional method to align a face is eye detection [3], [4]. When a face has already been detected, it is possible to

Manuscript received Feb. 16, 2010; revised Sept. 7, 2010; accepted Sept. 30, 2010.

This work was supported by the R&D program of the Korea Ministry of Knowledge and Economy (MKE) and the Korea Evaluation Institute of Industrial Technology (KEIT) [Project KI001813, Development of HRI Solutions and Core Chipsets for u-Robot].

Kyu-Dae Ban (phone: +82 10 5000 7357, email: kdban@hanmail.net) and Yun Koo Chung (email: ykchung@etri.re.kr) are with the IT Convergence Technology Research Laboratory, ETRI, Daejeon, Rep. of Korea, and also with the Department of Computer Software & Engineering, University of Science and Technology, Daejeon, Rep. of Korea.

Jaeyeon Lee (email: leejy@etri.re.kr), DoHyung Kim (email: dhkim008@etri.re.kr), and Jaehong Kim (email: jhkim504@etri.re.kr) are with the IT Convergence Technology Research Laboratory, ETRI, Daejeon, Rep. of Korea.

doi:10.4218/etrij.11.1510.0022

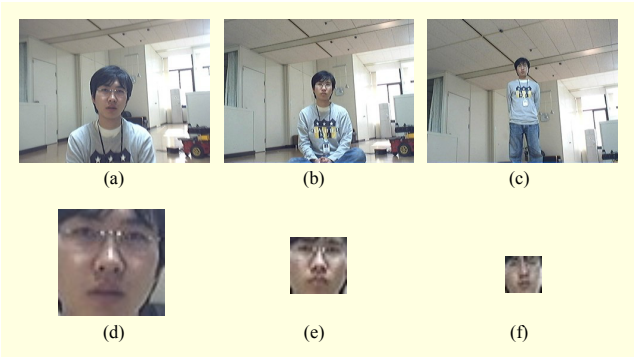

Fig. 1. Captured images at (a) 1 m, (b) 2 m, and (c) 3 m, and their respective cropped face images with face widths of (d) 53 pixels, (e) 28 pixels, and (f) 13 pixels. The resolution of the original image is 320×240, and horizontal FOV is 52 degrees.

normalize its size and skew by finding the darkest two points in the eye candidate regions. If the size of the face image is large (about 100×100 pixels and over), the pupils of the eyes can be detected easily, but this is difficult when the face size is small. In Fig. 1, the size of the face is only about 13×13 pixels in a 320×240 pixel image when captured at three meters away with a horizontal field of view (FOV) of 52 degrees. In this face image, the eyes are extremely blurred. Therefore, an eye-detection-based face alignment method may not be applicable to a robot environment.

Recent face alignment investigations [5]-[14], except for eye detection, are almost all focused exclusively on the active shape model [15], active appearance model (AAM) [16], or their variants. The performance of the AAM depends on modeling and fitting. The AAM is sensitive to initial parameter values and is apt to be trapped at the local minima in the fitting process. If the face image quality is extremely low, it is very difficult to deal with features accurately. Most recent face alignment researchers use high quality images that have clear edge information. Thus, face alignment with the AAM method is also not appropriate for robot environments where the quality of the image is poor and only a little edge information is available.

We assume a robot needs to recognize people looking into the robot's eyes, that is, the robot's camera. Therefore, our face alignment method is aimed to target at frontal faces. Considering a robot environment, we would like to focus attention on how to align a tiny and blurred face as a preprocessing step in face recognition. With such facial images, we assume that exact eye detection is difficult, even for humans. For this reason, we propose a new method of face alignment. In brief, our method is used to train the classifiers that divide aligned and non-aligned faces. This method simply uses more of the face region (including eyes, nose, and mouth components) than eye-detection-based face alignment. The most fascinating discovery is, in our own database, final face recognition rates under our modified census transform (MCT)-AdaBoost face alignment method are better than ground-truth-based face alignment at a distance.

II. Proposed Face Alignment Method

1. Face Alignment as Classification

To find an exact mugshot region in an unevenly detected face, we need a metric function for assessing how well the face is aligned. A solution for the alignment score can be found in the confidence value of Real-AdaBoost. Real-AdaBoost is a generalization of Discrete AdaBoost, which appears in the work of Freund and others [17]. It uses real-valued 'confidence-rated' outputs instead of {−1, 1} of discrete AdaBoost. The weak learner for this generalized boosting generates a mapping, $f_m(x): X \to R$, where X is the domain of the predictive feature x. In the prediction, the sign of $f_m(x)$ provides the classification, and its magnitude provides the measure of 'confidence' [18].

Eventually, we assumed that face alignment is a classification problem to separate well-aligned from non-aligned faces. In the following, we present the feature and training methods.

2. Feature Selection

In order to reflect the robustness toward illumination variation, MCT [19] features were selected. The equation of MCT is

$$\Gamma(x) = \bigotimes_{\mathbf{x}' \in N'(x)} \zeta(\bar{I}(x), I(x')), \quad (1)$$

where $\zeta(I(x), I(x'))$ indicates a comparison function, which is 1 if $I(x) < I(x')$ or 0; $\otimes$ denotes the concatenation operation; $N'(x)$ is a local spatial neighborhood of the pixel at x; and $I(x)$ and $\bar{I}(x)$ denote the pixel's gray intensity at x and the mean of neighborhood intensity, respectively. MCT can determine the $511(2^9-1)$ structure kernels defined on a 3×3 neighborhood. Figure 2 shows an example illustration of an MCT.

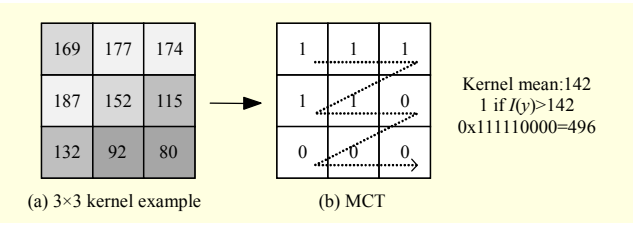

Fig. 2. Example of modified census transform.

3. Training Classifier

AdaBoost can be used to combine the feature selection and classifier. The training of AdaBoost usually shares a cascaded classifier for object detection; however, in the aspect of feature selection, we expected that many weak classifiers would significantly help to find a reliable confidence value. Thus, we trained only a single-stage classifier with many features. The number of trained locations for that classifier was 1,000 (the window can have a maximum of (45–2) × (40–2) = 1,634).

To train the classifier, we use the AdaBoost procedure of Fröba and others [19]. In this subsection, we briefly describe their boosting procedure.

The resulting values of the MCT feature are integer indices lying within the range of [0 to 511], and each index of a specific pixel location is used to construct the weak classifiers w_x. In brief, if it is more likely to show up in a positive class, the weak classifier is assigned 0 at kernel index γ; otherwise, it is assigned 1. The pixel classifier h_x is the weighted sum of all weak classifiers w_t at location x. Thus, this is comprised of a lookup table of length 511, which is the number of MCT kernel indices. The lookup table holds a weight for each kernel index. The final stage classifier $H(\Gamma)$ is the sum of all pixel classifiers h_x.

$$H(\Gamma) = \sum_x h_x(\Gamma(x)) . \quad (2)$$

Fröba and others use the decision rule, $H(\Gamma) \le T$, where T is the score threshold to classify a face or non-face. However, our face alignment problem is classifying an aligned face among a roughly detected face image. Therefore, we need to train a strong classifier that can divide aligned and non-aligned faces, which means that

$$H(\Gamma_i) < H(\Gamma_j) , \quad (3)$$

where $\Gamma_i \in A$ and $\Gamma_j \in N$, where A and N are the classes of aligned and non-aligned faces, respectively. We expected that (3) would be satisfactory if training sets containing positive and negative examples are well organized for modeling aligned and non-aligned faces. The construction of positive and negative training sets will be introduced in the next session.

4. Positive and Negative Training Samples

For training the classifiers of the proposed face alignment method, it was necessary to obtain a large database with aligned and non-aligned face images. We gathered a large number of face images (about 14,000 images) from the Internet. First, we marked eye position manually, and then we made aligned and non-aligned face images through the normalization process. More specifically, all aligned faces have to be normalized to have identical eye locations which, in our case, are (10, 15) and (30, 15) in a 40×45 pixel-sized face. On the other hand, non-aligned face images were generated from images in the positive samples with random modification of eye positions. In such ways, our positive image set contains 10,000 aligned face images, and our negative image set has 50,000 non-aligned face images. Figure 3 shows some sample images of positive and negative training sets for face alignment.

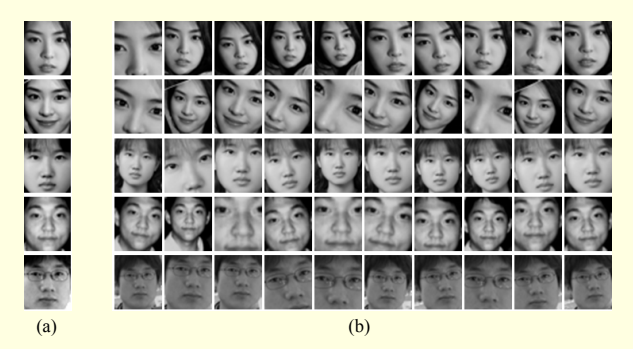

Fig. 3. Sample images for training AdaBoost face alignment: (a) positive and (b) negative sample images, where images in (b) are generated from (a).

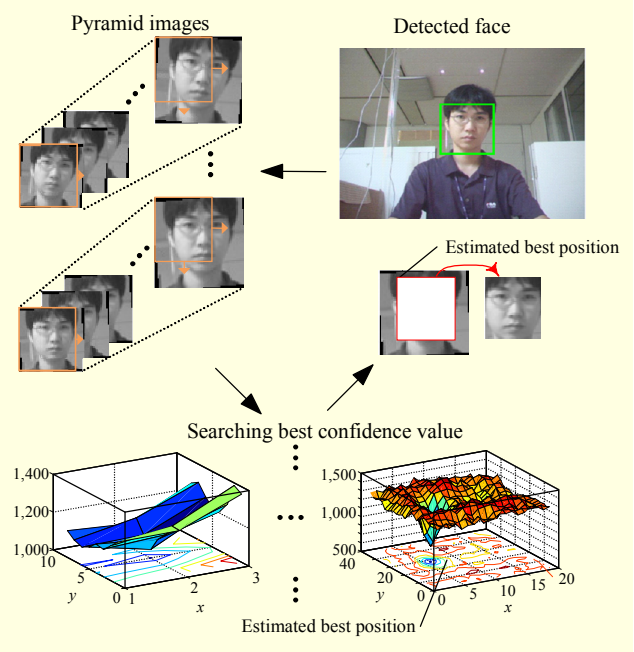

Fig. 4. Overview of MCT-AdaBoost face alignment.

5. Face Alignment

Figure 4 illustrates the proposed face alignment process. The process includes the construction of the search space, searching

for the maximum confidence value of the classifier, and cropping the mugshot region from the most appropriate image of the search space pool. In order to find a mugshot to recognize, an exact face region is searched from a rough face region. A rough face means that the face region is more expanded than that from face detection because a detected face region often does not include the entire face region that is needed for recognition. The size of a mugshot, which is used as a feature, is 40×45 pixels for our face recognizer. Thus, the search space was set to be bigger than 40×45 pixels. The size of each face window (mugshot) was fixed, so we reconstructed the search space by considering the size and rotation variations. We formed search spaces with 10 different sizes and 5 different angles. In one candidate face image of the search space pool, a distribution of alignment score is created using the confidence values, which are an output of the MCT-AdaBoost face alignment method, of the sliding windows. Among these alignment score distributions, an image having a minimum value (maximum confidence) is regarded as appropriate when the face is mugshot-sized and not rotated, that is, aligned. The position of the maximum confidence value is offset to crop an aligned face from that image.

III. Experiments and Results

1. Database

We collected a realistic and large database to evaluate the performance of the proposed face alignment method in a robot environment. This database is different from the database for training alignment engine. The database contains frontal-view face images of 10 users within three ranges of distance, 1 m, 2 m, and 3 m, during one month in a general home environment. The number of users was determined after consideration of the family members.

The process of building the database is described as follows. Our basic assumption is that users are looking at the robot in various face poses and talking. Users can laugh, occasionally grimace, and look at other places. Most of the time, however, they maintained eye contact while the images were captured. Due to the limitation of the camera FOV, we had to capture image sequences of user in different posture at different distances. For example, users can sit on the floor at 1 m, sit on a chair or stand at 2 m, and stand at 3 m. Variations in illumination are also included to reflect uncontrolled environments. The database contains a total of 20,750 images. Figure 5 shows some sample images at different distances. The resolution of the captured images is 320×240 pixels, and the horizontal FOV of the camera was 52 degrees.

We used 200 images for the gallery, which were captured from a 1 m distance during the first five days. For the probe images, 6850×3 pixel images are used, captured at different distances, 1 m, 2 m, or 3 m.

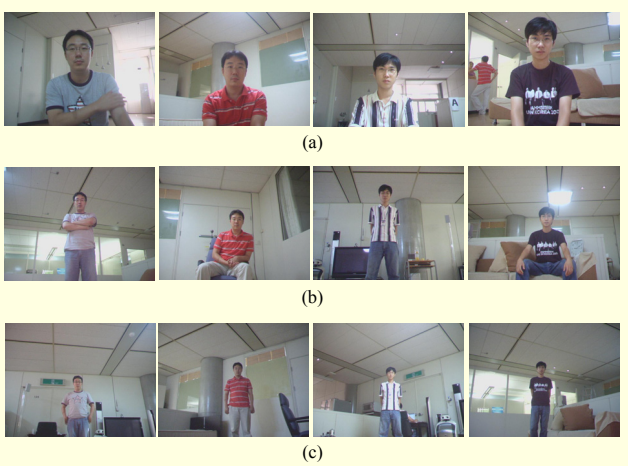

Fig. 5. Sample images of generated database at (a) 1 m, (b) 2 m, and (c) 3m distance.

To develop our face alignment and recognition algorithm, we needed to build a ground truth of our database, but the facial sizes are too small to create it exactly. Therefore, we marked the eye positions at the sub-pixel level by magnifying the face regions by about ten times.

2. Comparison of Alignment Methods

A face alignment result itself may not be sufficient for awareness of the significance of the face alignment method. For example, if we obtain a result with a difference of two pixels in left eye position between the face alignment method and ground truth, it is difficult to instinctively understand the worth of this value (two pixels). Therefore, we show the results of face recognition rates for the evaluation of face alignment. A comparison of changes in recognition rates is helpful to judge which face alignment engine is better and how much better it is. Note, of course, that comparison of alignment methods is only possible when the user's faces are detected already. Our face detector can detect a user's face when the eyes are visible. Ideally, every detected face was of a full frontal view. Figure 6 shows an abbreviated flow diagram of the comparison experiments used. For comparison, we implemented three alignment methods.

In the first method, we directly used a detected face instead of an aligned face. However, the detected face was normalized in its size and skew for recognition. Here, we need to describe our face detector further. For detection of a small face, we trained the weak classifiers of AdaBoost with small frontal faces (16×16 pixels).

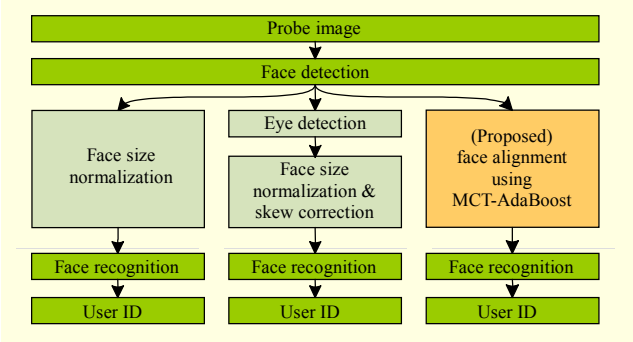

Fig. 6. Flowchart for face alignment comparison. Evaluation of face alignment is done through face recognition.

The number of training samples is 18,000 for positive images and 60,000 for negative ones. The negative samples were cropped and extracted into 5,000 high-resolution non-face images. We also used four stages for cascade training. Details of the training algorithm are provided in [20]. This face detector can detect a human face at 3 m away from a robot, assuming that the robot camera has a 52 degree FOV and a resolution of 320×240 pixels.

The second and third methods use eye detection. We adopt two eye detection methods. The first is the AdaBoost method. For eye detection, the fast object classification approach of Viola and others [21] is attracting significant attention. Based on their approach, many different authors have made classifiers for public use. Castrillón Santana and others [22] analyzed the individual performance of these public classifiers. They showed that their eye detector has the best results in their facial feature detection experiments, especially in eye detection. The number of stages is 16 for the left eye and 18 for the right, and the classifier in the cascade is 18×16 pixels.

The second eye detection method is based on the method of Yoon and others [23] because of its computational efficiency and simple structure. The process is as follows: The first step is adaptive sobel edge detection of a face image. The second step is a two-pass labeling algorithm. The last step is the verification of the size, shape, and symmetry using face model knowledge for eye detection. For convenience, we call this method an edge eye detector.

Comparison of face alignment ability is done through face recognition rates. Our face recognizer is based on a composite of multiple features and matching algorithms. We adopted a multiple principal component analysis (PCA) and edge distribution methods, as features for human face representations. These features are projected onto a new intra-person/extra-person similarity space that is comprised of several similarity measures. The final evaluation is done using a support vector machine (SVM). Further details can be found in [24].

3. Experimental Results

Figure 7 and Table 1 show the face recognition rates according to the various face alignment methods. The face recognition results show that proposed AdaBoost with the MCT feature is better than eye-detector-based face alignment and face detector alone. The proposed method shows an even higher rate than ground-truth-based face alignment within all distance ranges. This result is reasonable under the assumption that manually marking the exact eye or iris region is difficult in a small and blurred face image. In the following discussion, we will address this in depth.

Table 2 shows the average processing time according to the

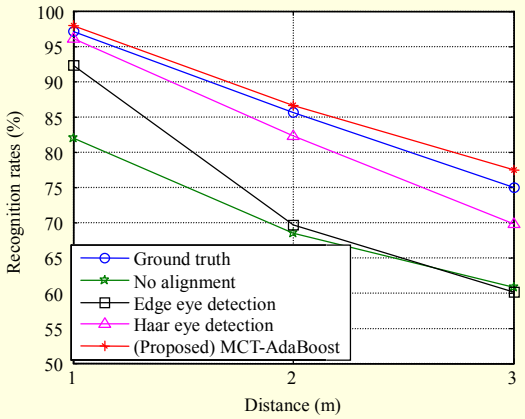

Fig. 7. Face recognition rates according to alignment methods.

Table 1. Face recognition rates according to alignment method.

Alignment method	Recognition rate (%) at distance		
	1 m	2 m	3 m
Ground truth	97.01	85.57	74.86
No alignment (face detection only)	82.00	68.34	60.72
Edge eye detector	92.25	69.63	60.02
Haar eye detector	96.05	82.23	69.74
(Proposed) MCT-AdaBoost	97.07	86.56	77.36

Table 2. Processing time according to alignment method.

Alignment method	Processing time (ms)
Ground truth	14
No alignment (face detection only)	53
Edge eye detector	54
Haar eye detector	54
(Proposed) MCT-AdaBoost	118

face alignment method used. This time includes face detection and face alignment. The system used in this experiment is a 2.6 GHz Pentium 4. The software used in this work has been implemented in C++. No GPU-based acceleration was used. We used the Intel OpenCV library [25] only for the Haar eye detector.

4. Discussion

In the general face recognition problem, it is unusual that a recognition rate using the ground truth is lower than that using other algorithms. However, this is possible when the target of manual marking is not an ID but the eye location in a blurred face image. In the case of a low-quality blurred face image obtained in a robot environment, the manual marking of the eye location is difficult, even if the images are resized through high magnification, because most of the face components are significantly blurred.

Figure 8(b) shows a simple example of a resized image that was cropped from a detected face at 3 m and 640×480 resolution in image of Fig. 8(a). Note that Fig. 8(a) has higher resolution than that of face recognition databases. The original face size was 36×48 pixels, and the resized face is 360×480. Since this image is magnified ten times, the process of eye-marking makes a sub-pixel level ground truth. Ten eye positions were marked by ten people. The mean position of left eye marks is (123.9, 121.3), and its standard deviation is (2.6, 6.1) pixels. According to the research of Wang and others [13], if we assume that the exact eye position is equal to the mean, this roughly 6 pixel eye location error can reduce the face recognition accuracy by over 10%. This implies that facial recognition using face alignment with a manual marking of the eye position cannot achieve the highest rates.

As can be seen in Fig. 7 and Table 1, the recognition rates of 1 m reveal that conventional face-alignment-based eye detection is applicable when the distance between a robot and the user is short. When the distance is long, however, it is difficult to detect the exact eye position. The recognition rates of eye-detection-based methods are proportionally lower with a longer distance. Because the distance is longer, the detailed information around the eyes is loose. On the other hand, the proposed method uses the entire appearance of the face. This means that our alignment method utilizes more information than those based solely on eye detection. Therefore, the recognition rate of our method can be higher than other methods at a distance.

Even though the processing time of our proposed algorithm is about two-times slower than the eye detection methods, 118 ms (over 8 frames/s) can be acceptable in robot applications.

In order to align a face, the feature used in this study is MCT.

(a)

(b)

Fig. 8. (a) Original 320×240 image and (b) its face magnification. Note that making ground truth is difficult because most face components are significantly degraded. Also, center points of eyes are different between individuals.

However, this would not exclude the use of other features such as Haar-like features [21]. A more important point to be emphasized is that the negative training samples were non-aligned faces, instead of arbitrary non-facial images normally used in the training of face detectors. By using non-aligned faces, we could discriminate the best aligned faces from non-aligned faces efficiently.

IV. Conclusion

In this paper, face alignment using the MCT-AdaBoost technique is proposed to align a detected face image. We collected a large database in a real home environment to assess the performance of several face alignment methods. Even though conventional eye-detection-based face alignment gave a low face recognition rate at a distance, the proposed MCT-AdaBoost-based face alignment method gave a higher recognition rate than manual face alignment based on ground truth. The proposed face alignment method can be applied to various applications dealing with a highly degraded image, such as that from an intelligent service robot, long-range surveillance, and so on. Future work will focus on speed improvement of the proposed algorithm using optimization algorithms such as the Gradient descent method.

References

[1] L. Aryananda, "Recognizing and Remembering Individuals: Online and Unsupervised Face Recognition for Humanoid Robot," *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sept. 2002, pp. 1202-1207.

[2] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cog. Neuroscience*, vol. 3, no. 1, 1991, pp. 71-86.

[3] D. Hansen and Q. Ji, "In the Eye of the Beholder: A Survey of Models for Eyes and Gaze," *IEEE Trans. Pattern Anal. Mach. Intell.* (submitted).

[4] H. Kim and W. Kim, "Eye Detection in Facial Images Using Zernike Moments with SVM," *ETRI J.*, vol. 30, no. 2, Apr. 2008, pp. 335-337.

[5] M. Zhao and T. Chua, "Markovian Mixture Face Recognition with Discriminative Face Alignment," *IEEE Int. Conf. Autom. Face Gesture Recog.*, Sept. 2008, pp. 1-6.

[6] L. Wang, X. Ding, and C. Fang, "Generic Face Alignment using an Improved Active Shape Model," *Int. Conf. Audio, Language, Image Process.*, July 2008, pp. 317-321.

[7] F. Kahraman and M. Gokmen, "Illumination Invariant Three-Stage Approach for Face Alignment," *IEEE Intl. Conf. Image Process.*, Oct. 2006, pp. 2073-2076.

[8] Y. Su, H. Ai, and S. Lao, "Real-Time Face Alignment with Tracking in Video," *IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1632-1635.

[9] Y. Zhou et al., "A Bayesian Mixture Model for Multi-View Face Alignment," *IEEE Computer Society Conf. Computer Vision Pattern Recog.*, vol. 2, June 2005, pp. 741-746.

[10] K. Choi, J. Ahn, and H. Byun, "Face Alignment Using Segmentation and a Combined AAM in a PTZ Camera," *Int. Conf. Pattern Recog.*, vol. 3, Sept. 2006, pp. 1191-1194.

[11] X. Liu, "Generic Face Alignment Using Boosted Appearance Model," *IEEE Conf. Computer Vision Pattern Recog.*, June 2007, pp. 1-8.

[12] Y. Huang et al., "Face Alignment under Variable Illumination," *IEEE Int. Conf. Autom. Face Gesture Recog.*, May 2004, pp. 85-90.

[13] P. Wang, L. Tran, and Q. Ji, "Improving Face Recognition by Online Image Alignment," *Int. Conf. Pattern Recog.*, vol. 1, Sept. 2006, pp. 311-314.

[14] F. Kahraman, B. Kurt, and M. Gokmen, "Robust Face Alignment for Illumination and Pose Invariant Face Recognition," *IEEE Conf. Computer Vision Pattern Recog.*, June 2007, pp. 1-7.

[15] P. Wang et al., "Automatic Eye Detection and its Validation," *IEEE Computer Soc. Conf. Computer Vision Pattern Recog.*, June 2005, pp. 164-164.

[16] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, June 2001, pp. 681-685.

[17] Y. Freund and R.E. Shapire, "A Short Introduction to Boosting," *J. Japanese Society for Art. Intell.*, no. 14, Sept. 1999, pp. 771-780.

[18] J. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting," *Annals of Statistics*, vol. 38, no. 2, Apr. 2000, pp. 337-374.

[19] B. Fröba and A. Ernst, "Face Detection with the Modified Census Transform," *Int. Conf. Autom. Face Gesture Recog.*, 2004, pp. 91-96.

[20] B. Jun and D. Kim, "Robust Real-Time Face Detection Using Face Certainty Map," *Intl. Conf. Biometrics*, vol. 4642, no. 125, Aug. 2007, pp. 29-38.

[21] P. Viola and M. Jones, "Robust Real Time Object Detection," *IEEE ICCV Workshop Statistical Computational Theories of Vision*, July 2001, pp. 1-25.

[22] M. Castrillón Santana et al., "Face and Face Feature Detection Evaluation - Performance Evaluation of Public Domain Haar Detectors for Face and Face Feature Detection," *VISAPP*. vol. 2, 2008, pp. 167-172.

[23] Ho-Sub Yoon et al., "Face Component Detection on the Natural Office Scene," *Intl. Conf. Adv. Intell. Mechatronics*, 1997, p. 28.

[24] D. Kim et al., "A Non-Cooperative User Authentication System in Robot Environments," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, May 2007, pp. 804-811.

[25] Intel. Open Source Computer Vision Library (OpenCV). Available at: http://sourceforge.net/projects/opencvlibrary/

Kyu-Dae Ban received the BS in electrical and computing engineering from Chungbuk National University in 2005. He is currently working towards the PhD in the Department of Computer & Software Engineering at the University of Science and Technology, Rep. of Korea. His research interests are HRI and face recognition.

Jaeyeon Lee received his PhD from Tokai University, Japan, in 1996. He has been a research scientist at ETRI since 1986. His research interests include robotics, pattern recognition, and computer vision.

DoHyung Kim received the PhD from Pusan National University in 2009. He has been a research scientist at ETRI since 2002. His research interests include human robot interaction, computer vision, and pattern recognition.

Jaehong Kim received his PhD from Kyungpook National University in 1996. He is a research scientist at ETRI. He is interested in the silver-care aspects of human-robot interaction.

Yun Koo Chung received his PhD from the Wayne State University in 1991. He is a research scientist at ETRI. His research interests include image processing and computer vision.