Gaze Estimation from Low Resolution Images

Yasuhiro Ono, Takahiro Okabe, and Yoichi Sato

Institute of Industrial Science, The University of Tokyo,
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
{onoy,takahiro,ysato}@iis.u-tokyo.ac.jp
http://www.hci.iis.u-tokyo.ac.jp/~onoy/

Abstract. The purpose of this study is to develop an appearance-based method for estimating gaze directions from low resolution images. The problem of estimating directions using low resolution images is that the position of an eye region cannot be determined accurately. In this work, we introduce two key ideas to cope with the problem: incorporating training images of eye regions with artificially added positioning errors, and separating the factor of gaze variation from that of positioning error based on N-mode SVD (Singular Value Decomposition). We show that estimation of gaze direction in this framework is formulated as a bilinear problem that is then solved by alternatively minimizing a bilinear cost function with respect to gaze direction and position of the eye region. In this paper, we describe the details of our proposed method and show experimental results that demonstrate the merits of our method.

Key words: gaze estimation, low resolution, appearance-based method, positioning, N-mode SVD

1 Introduction

Gaze direction is an important cue for understanding human activities since they are considered to be well correlated with our focus of attention. Thus robust and nonintrusive estimation of gaze direction, hereafter referred to as *gaze estimation*, can be used effectively for a wide variety of applications. For instance, gaze estimation techniques can be used for determining how often and which part of a billboard is being looked at in a public space such as a shopping mall.

One of the key challenges for gaze estimation for such applications is that it is not always possible to capture high resolution images due to limitation of camera placement. Therefore, it is important to have techniques for gaze estimation from *low resolution images*. For example, consider the case of estimating gaze directions by using images captured by a surveillance camera already installed in an environment. It is likely that the camera is far from a subject, and thus only low resolution images of the subject's face are available.

Previously proposed methods for gaze estimation, which are classified into two approaches: *model-based methods* and *appearance-based methods*, are not suitable for the purpose of gaze estimation from low resolution images for several reasons.

Model-based methods usually require high resolution images of human faces to estimate gaze direction accurately. This is because gaze directions are determined from the eye's geometric features localized in images. Among model-based methods, the most commonly used techniques are the ones based on pupil corneal reflection [2, 4, 7, 17, 18]. A gaze direction is determined from the relative position of the pupil center

Fig. 1. Images of an eye with (a) high and (b) low resolutions. It is not trivial to accurately extract geometric features of the eye and feature points for positioning from the low resolution image.

and a glint reflected on the cornea of an eyeball. Other techniques use the position of an iris center or a pupil center obtained from edge detection or ellipse fitting for estimating gaze direction [5, 6, 14]. As we see in Figure 1, it is not trivial to extract the above features from low resolution images of an eye. In contrast, appearance-based methods can be used for estimating gaze directions from low resolution images because these methods use pixel values of eye regions for estimating gaze directions, and therefore it is not necessary to find the eye's geometric features in input images. Unlike model-based methods, appearance-based methods have had very few studies devoted to them. Some researchers have proposed gaze estimation methods using a neural network that is trained with eye images of known gaze directions [1, 10, 15]. Recently, Tan *et al.* developed a method based on nearest neighbor search that essentially looks for the nearest training image for a given input image in order to determine gaze direction for the input image [11].

Unfortunately, the previously proposed appearance-based methods share a common problem. They require eye regions to be accurately positioned in input images, which is not always easy due to the nature of low resolution images as we see in Figure 1. Even a slight *positioning error*, *i.e.*, error in the position of a cropped eye region, can degrade the accuracy of gaze estimation significantly. This important problem has not been addressed in the previous studies.

In this work, we introduce two key ideas in order to cope with the problem. One is to incorporate training images of eye regions with artificially added positioning errors. The other is to separate the factor of gaze variation from that of positioning error based on N-mode SVD (Singular Value Decomposition), which was recently introduced to the computer vision community by Vasilescu *et al.* [12]. We show that estimation of gaze direction in this framework is formulated as a bilinear problem that is then solved by alternatively minimizing a cost function with respect to gaze direction and the eye region's positioning. In order to examine how well the effect of positioning errors is removed by our method, we compared our method with an appearance-based method using PCA (Principal Component Analysis). As a result, we found that our method is able to estimate gaze directions with significantly higher accuracy than the PCA-based method.

The rest of this paper is organized as follows. In Section 2, we explain our proposed method for estimating gaze directions from low resolution images. In Section 3, we show experimental results demonstrating the merits of our proposed method. Finally, we present our concluding remarks in Section 4

2 Proposed Method

2.1 Overview

The appearance of an eye depends not only upon gaze direction but also upon identities of subjects, poses of a head, and imaging conditions such as image resolution, response of a camera, and illumination conditions. In the present study, we focus on the problem

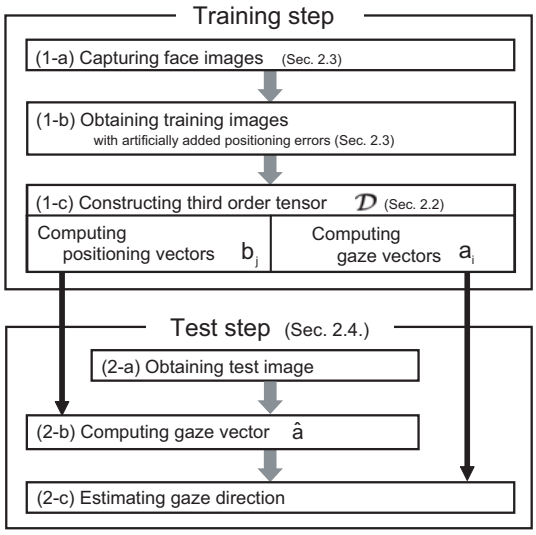

Fig. 2. Flowchart of our proposed method.

of estimating gaze direction from low resolution images of an eye. Therefore, we do not consider appearance variations due to other factors such as identities of subjects and poses of a head. We will describe our plan for future study to deal with those factors in the Conclusions section of this paper.

Our proposed method consists of two steps: *the training step* and *the test step* as summarized in Figure 2. In the training step, we first capture face images with different gaze directions (1-a), and obtain an enlarged set of training images of eye regions with artificially added positioning errors (1-b). Appearance variations due to gaze direction and positioning are then modeled based on 3-mode SVD (1-c). More specifically, we construct a third order tensor from the training images, and compute a pair of two feature vectors describing gaze direction and positioning, which we call *a gaze vector* and *a positioning vector* respectively, for each training image. In the test step, we extract the gaze vector of a test image (2-b), and finally estimate the gaze direction by comparing the extracted gaze vector with those of the training images (2-c). In Section 2.2, we explain how the appearance variation of an eye for different factors is modeled by using 3-mode SVD. Then, in Sections 2.3 and 2.4, we describe the training step and the test step of our proposed method in detail.

2.2 Appearance Modeling using 3-mode SVD

N-mode SVD is one of the natural extensions of ordinary (2-mode) SVD to multiple modes. Here, a *mode* is a factor affecting data. For instance, identities, viewpoints, illumination conditions, facial expressions, and also image pixels can be considered as modes in face recognition [12].

Our proposed method models variations in the eye region's appearance related to three factors, *i.e.*, gaze directions, positionings, and image pixels, by using *3-mode SVD*. Here, positioning means how eye regions are cropped in input images. Let us consider a set of images of the eye region with different gaze directions and positionings. We represent those images by using a third order tensor $\mathcal{D}$. Here, the component

$\mathcal{D}_{ijk}$ $(1 \leq i \leq I, 1 \leq j \leq J, 1 \leq k \leq K)$ of the tensor is the k-th pixel value in the image with the i-th gaze direction and the j-th positioning. I, J, and K are the total numbers of different gaze directions, different positionings, and image pixels.

We can represent the third order tensor as

$$\mathcal{D}_{ijk} = \sum_{l=1}^{I} \sum_{m=1}^{J} \sum_{n=1}^{K} \mathcal{Z}_{lmn} \left(U_{\mathrm{G}}\right)_{il} \left(U_{\mathrm{POS}}\right)_{jm} \left(U_{\mathrm{PIX}}\right)_{kn}, \tag{1}$$

where $\boldsymbol{U}_{\mathrm{G}} \in \Re^{I \times I}$, $\boldsymbol{U}_{\mathrm{POS}} \in \Re^{J \times J}$, and $\boldsymbol{U}_{\mathrm{PIX}} \in \Re^{K \times K}$ are basis matrices for gaze mode, positioning mode, and pixel mode respectively, and $\mathcal{Z}_{lmn}$, which represents interaction among basis matrices, is called the *core tensor*. Basically, they correspond to two orthonormal matrices and one diagonal matrix in ordinary SVD. We also denote Eq. (1) as

$$\boldsymbol{\mathcal{D}} = \boldsymbol{\mathcal{Z}} \times_1 \boldsymbol{U}_{\mathrm{G}} \times_2 \boldsymbol{U}_{\mathrm{POS}} \times_3 \boldsymbol{U}_{\mathrm{PIX}}. \tag{2}$$

The basis matrix for each mode is computed as follows. First, we unfold the tensor $\boldsymbol{\mathcal{D}}$ and construct a matrix. For instance, we unfold the tensor with respect to the gaze mode G to obtain the matrix $\boldsymbol{D}_{\mathrm{G}} \in \Re^{I \times KJ}$ as $\boldsymbol{D}_{\mathrm{G}} = \left[\boldsymbol{F}_1 \boldsymbol{F}_2 ... \boldsymbol{F}_K\right]$. Here, the matrix $\boldsymbol{F}_k \in \Re^{I \times J}$ $(1 \leq k \leq K)$ is a slice of the tensor $\boldsymbol{\mathcal{D}}$ with a fixed value of k. Then, by applying SVD to the matrix $\boldsymbol{D}_{\mathrm{G}}$ as $\boldsymbol{D}_{\mathrm{G}} = \boldsymbol{U}_{\mathrm{G}} \boldsymbol{\Sigma}_{\mathrm{G}} \boldsymbol{V}_{\mathrm{G}}^{\top}$, we obtain the basis matrix $\boldsymbol{U}_{\mathrm{G}} \in \Re^{I \times I}$ of the gaze mode. The basis matrices $\boldsymbol{U}_{\mathrm{POS}}$ for the positioning mode and $\boldsymbol{U}_{\mathrm{PIX}}$ for the pixel mode are computed similarly.

The core tensor $\boldsymbol{\mathcal{Z}}$ in Eq. (2) is computed by using the tensor $\boldsymbol{\mathcal{D}}$ and basis matrices $\boldsymbol{U}_{\mathrm{G}}$, $\boldsymbol{U}_{\mathrm{POS}}$, and $\boldsymbol{U}_{\mathrm{PIX}}$ as $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{D}} \times_1 \boldsymbol{U}_{\mathrm{G}}^{\top} \times_2 \boldsymbol{U}_{\mathrm{POS}}^{\top} \times_3 \boldsymbol{U}_{\mathrm{PIX}}^{\top}$.

We define gaze vectors $\boldsymbol{a}_i \in \Re^I$ $(1 \leq i \leq I)$ and positioning vectors $\boldsymbol{b}_j \in \Re^J$ $(1 \leq j \leq J)$ as

$$\left[\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_I\right] \stackrel{\mathrm{def}}{=} \boldsymbol{U}_{\mathrm{G}}^{\top}, \quad \left[\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_J\right] \stackrel{\mathrm{def}}{=} \boldsymbol{U}_{\mathrm{POS}}^{\top}. \tag{3}$$

In other words, $\boldsymbol{a}_i \in \Re^I$, for example, represents the i-th gaze direction in the feature space of gaze direction.

For each pair of the gaze vector $\boldsymbol{a}$ and the positioning vector $\boldsymbol{b}$, the image $\boldsymbol{d}$ of a corresponding eye region is given by using a third order tensor $\boldsymbol{\mathcal{B}}$

$$\mathcal{B}_{ijk} \stackrel{\mathrm{def}}{=} \sum_{l=1}^{K} \mathcal{Z}_{ijl} \left(U_{\mathrm{PIX}}\right)_{kl} \tag{4}$$

as

$$d_k = \sum_{i=1}^{I} \sum_{j=1}^{J} \mathcal{B}_{ijk} a_i b_j = \sum_{i=1}^{I} \sum_{j=1}^{J} B_{k(ij)} a_i b_j. \tag{5}$$

Here, we represent the tensor $\mathcal{B}$ in the matrix form B. $B_{k(ij)}$ is the $(k, I \times (j-1) + i)$ component of the matrix B_{PIX}, which is obtained by unfolding $\boldsymbol{\mathcal{B}}$ with respect to the pixel mode.

2.3 Training Step

In the training step, variations in appearance of eye regions are learned from training images that are down-sampled from high resolution images of eyes. This is done because, unlike test images, high resolution training images are easily available, and, more importantly, the position of an eye can be found accurately in high resolution images by using existing techniques. For the present study, we used our feature-based face tracker [8] for finding eye corners.

We first capture a set of high resolution images of an eye with different but known gaze directions. Then, eye region images without positioning errors are obtained by cropping rectangular regions from down-sampled images by using the positions of eye corners. In addition, eye region images with artificially added positioning errors are obtained by moving the positions of eye corners.

After a set of training images of eye regions is created, we construct the third order tensor $\mathcal{D}$ from the training images and obtain the gaze vectors $\boldsymbol{a}_i$ $(1 \leq i \leq I)$ and the positioning vectors $\boldsymbol{b}_j$ $(1 \leq j \leq J)$ from Eq. (3) as described in Section 2.2. Additionally, we prepare the matrix $B_{k(ij)}$ from Eq. (4).

2.4 Test Step

For each test image, an eye region is found first. It should be noted that, unlike in the training step, the image resolution of the eye region is not necessarily high. However, some existing techniques for facial component detection, *e.g.*, the AdaBoost algorithm [3] and the Gabor-like feature filtering scheme [16], can find eye regions even in low resolution test images.

After an eye region is found, the gaze vector is computed for the eye region. Then, the gaze direction for the test image is determined from the gaze vector. We will explain each of the steps in this section.

(1) Extraction of Gaze Vectors In order to extract feature vectors from test images, two methods based on projections have been proposed. The first one proposed by Vasilescu [12] for face recognition projects a test image into the feature space of face identity by using a set of matrices. The method uses one matrix per each combination of indices except for that corresponding to identity mode, and thus only yields a set of candidates for the correct feature vector.

The second method recently proposed by Vasilescu [13] uses a single matrix independent of specific values of modes. This method is more elegant than the first one in the respect that it can simultaneously extract unique feature vectors of all modes from a test image. However, the method is not applicable to our purpose of extracting feature vectors from low resolution images. The method assumes that the number of pixels is larger than the product of the number of indices in each mode. This assumption, which requires $K \geq IJ$ in our case, is not satisfied.

Accordingly, we introduce an algorithm for extracting feature vectors from low resolution images in the context of 3-mode SVD. Let us consider a test image $\hat{\boldsymbol{d}}$, and an image constructed from a gaze vector $\boldsymbol{a}'$ and a positioning vector $\boldsymbol{b}'$ through Eq. (5). Then, we define a cost function $f(\boldsymbol{a}', \boldsymbol{b}')$ by $f(\boldsymbol{a}', \boldsymbol{b}') \stackrel{\text{def}}{=} \sum_{k=1}^{K} \left(\hat{d}_k - \sum_{i=1}^{I} \sum_{j=1}^{J} \left(B_{k(ij)} a'_i b'_j \right) \right)^2$, and estimate the feature vectors of the test image $(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}})$ by minimizing the cost function as $(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = \arg\min_{\boldsymbol{a}' \in \Re^I, \boldsymbol{b}' \in \Re^J} f(\boldsymbol{a}', \boldsymbol{b}')$.

This cost function is bilinear, that is, it is linear with respect to one variable $\boldsymbol{a}'$ when the other variable $\boldsymbol{b}'$ is fixed, and vice versa. Therefore, we can directly and uniquely extract the feature vectors by alternatively minimizing the bilinear cost function in a similar manner to Shum [9]. Our proposed method thus relaxes the requirement with respect to the number of pixels from $K \geq IJ$ to $K \geq (I + J)$. In other words, for test images with a fixed image resolution, our method can use a wider variety of training images than the previously proposed method [13].

More specifically, the solutions of linear problems with respect to one variable $\partial f/\partial a'_i = 0 \ (1 \leq i \leq I)$ and $\partial f/\partial b'_j = 0 \ (1 \leq j \leq J)$ result in

$$\boldsymbol{a}' = \boldsymbol{M}^+ \hat{\boldsymbol{d}}, \quad \boldsymbol{b}' = \boldsymbol{N}^+ \hat{\boldsymbol{d}}, \quad (M)_{ki} \stackrel{\text{def}}{=} \sum_{j=1}^{J} B_{k(ij)} b'_j, \quad (N)_{kj} \stackrel{\text{def}}{=} \sum_{i=1}^{I} B_{k(ij)} a'_i, \quad (6)$$

where the matrix $\boldsymbol{M}^+$ is the pseudoinverse of $\boldsymbol{M}$. Therefore, we can assign, for example, the initial value of $\boldsymbol{b}'$ to $\boldsymbol{b}'^{(0)}$, and alternatively update the feature vectors according to Eq. (6) until they converge. Actually, we terminate the iteration when $\Delta f(n) \stackrel{\text{def}}{=} f(\boldsymbol{a}'^{(n)}, \boldsymbol{b}'^{(n)}) - f(\boldsymbol{a}'^{(n-1)}, \boldsymbol{b}'^{(n-1)})$ is less than the predefined threshold. Here, we denote the feature vectors at the n-th iteration as $\boldsymbol{a}'^{(n)}$ and $\boldsymbol{b}'^{(n)}$. In this way, the gaze vector is determined up to an unknown scale factor. Therefore, we normalize $\boldsymbol{a}'$ by using the L_2 norm to obtain the gaze vector $\hat{\boldsymbol{a}}$ for the given test image.

In the current implementation, we choose the initial value of $\boldsymbol{b}'$ according to $(\boldsymbol{a}'^{(0)}, \boldsymbol{b}'^{(0)}) = \arg\min_{\boldsymbol{a}' \in \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_I\}, \boldsymbol{b}' \in \{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_J\}} f(\boldsymbol{a}', \boldsymbol{b}')$. Namely, we search for the combination of the gaze and positioning vectors of training images that yields the image most similar to the test image in the least-square sense. Though the combination of the initial value might provide local minima of the cost function, our experimental results imply that the local minima do not affect the results seriously.

(2) Estimation of Gaze Direction Finally, we determine the gaze direction for the given test image by using the obtained gaze vector. We find three gaze directions of the training images nearest to that of the test image, and calculate the gaze direction of the test image by interpolating them. We do this because we need at least three gaze directions to represent an arbitrary gaze direction by interpolation. First, we find the index of the gaze vector of the training image that is the closest to the obtained gaze vector as $i\,(1) = \arg\min_{i \in \{1,2,\ldots,I\}} |\hat{\boldsymbol{a}} - \boldsymbol{a}_i|^2$. Similarly, we find the indices $i(2)$ and $i(3)$ of the second and third closest gaze vectors. Then we determine the gaze direction by interpolating the three gaze directions such that the interpolated gaze vector becomes the closest to the obtained gaze vector of the test image. This is done by choosing the three weights $w_p \ (p = 1, 2, 3)$ that minimize $|\hat{\boldsymbol{a}} - \sum_{p=1}^{3} w_p \boldsymbol{a}_{i(p)}|^2$ subject to $0 \leq w_p \leq 1$ and $\sum_{p=1}^{3} w_p = 1$. Then the gaze direction $\boldsymbol{g}$ is given as $\boldsymbol{g} = \sum_{p=1}^{3} w_p \boldsymbol{g}(p)$, where $\boldsymbol{g}(p)$ is the gaze direction of $i(p)$.

3 Experiments

3.1 Eye Images for Experiments

(1) Imaging Conditions We captured facial images of five individuals, and estimated gaze direction of each subject using our proposed method. In our experiments, we evaluated the accuracy of gaze estimation for each subject separately, *i.e.*, using training and test images of the same subject for gaze estimation. This was done because we do not deal with appearance variation due to different identities of subjects. To quantitatively evaluate the accuracy of our method, we captured images while subjects stared at targets appearing on an 18-inch SXGA monitor placed at a distance of 50cm from the subject's face. Since we calibrated the relative position of the monitor in advance of capturing images, we could calculate gaze direction corresponding to a 2D position on a monitor when a user was looking at the position.

Figure 3 shows the target positions displayed on the monitor: circles for training and crosses for test. Twenty training images and 32 test images were taken for each subject.

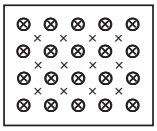

Fig. 3. A layout of targets displayed on the surface on an LCD monitor: circles for training and crosses for test.

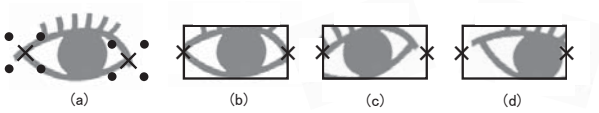

Fig. 4. (a) A schematic illustration of the correct corners of an eye (crosses) and the points used for artificially representing incorrect positioning (dots). (b) An illustration of an eye image cropped based on the correct positioning. (c) and (d) show those cropped with the incorrect positionings. We cropped eye regions so that the feature points in Figure 4 (a) are aligned to the crosses on both sides.

Since we do not consider face pose change in this study, we asked subjects not to move their heads while images were being taken. Each subject was asked to move a mouse pointer to a randomly appearing target and press the mouse button while the pointer was placed on the target to capture a face image of the subject staring at the target.

(2) Cropping Eye Regions After all face images were captured, we prepared eye region images for both training and testing. Note that we used down-sampled images for test images instead of images captured at low resolution in our experiments. This was necessary for evaluating the accuracy of our method quantitatively. The use of down-sampled test images enables us to investigate (i) how the accuracy of gaze estimation is affected by inaccurate positioning of eye regions, and (ii) how the estimation accuracy changes depending on the resolution of test images.

Eye regions were cropped from down-sampled images by using positions of eye corners found by our feature-based face tracker [8] and additional positions that were shifted diagonally from those true eye corners by one step[1]. Figure 4 (a) shows a schematic illustration of true eye corners and additional positions with artificially added positioning errors. In this way, we prepared 25 ($= 5 \times 5$) eye region images per gaze direction. All the images were aligned by affine transformations as illustrated in Figure 4 (b), (c), and (d).

For testing our method with different image resolutions, we used images with 16×48, 8×24, and 4×12 pixels as shown in Figure 5. As we see in those figures, it is impossible to localize geometric features such as the iris and cornea of an eye if image resolution is too low.

We show examples of eye regions for different gaze directions in Figure 6 (a). Those regions were cropped by using correct positions of eye corners. We also show eye regions for the same gaze direction but cropped with positioning errors in Figure 6 (b). From these examples, we see it is not trivial to estimate gaze directions from low resolution images without being affected by poor positioning accuracy.

[1] We define one step as 4 pixels, 2 pixels, or 1 pixel for each eye image with 16×48, 8×24, or 4×12 pixels respectively.

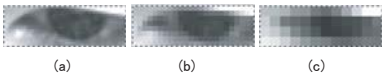

Fig. 5. Example images of an eye with (a) 16×48, (b) 8×24, and (c) 4×12 pixels.

Fig. 6. (a) Images cropped based on the correct positioning with various gaze directions. (b) Images cropped based on the various positionings with a fixed gaze direction.

3.2 Experimental Results

We quantitatively evaluated the performance of our proposed method, and compared the performance with that of a method based on conventional PCA from three aspects: how the estimation accuracy changes depending on image resolution, positionings, and individuals. The PCA-based method does not treat variations due to changes in gaze direction and position separately, and projects a test image into the feature space defined by the principal axes computed by using all training images with various gaze directions and positionings. The feature vector of one gaze direction is defined by the average of feature vectors computed for images with the same gaze direction but various positionings. In order to alleviate any bias due to brightness variation among images, we normalized training and test images for both our method and the PCA-based method so that pixel values in each image have zero mean and unit variance. We used 3-dimensional feature space for both our method and the PCA-based method to estimate gaze direction.

Estimation Error against Image Resolution First, we show errors of gaze estimation against image resolutions in Figure 7 (a). The horizontal axis indicates the number of pixels in the eye images ($4 \times 12 = 48$, $8 \times 24 = 192$, and $16 \times 48 = 768$), and the vertical axis represents the average and standard deviation of errors over five subjects. This figure shows that the accuracy of our proposed method is higher than that of the PCA-based method. Hereafter, we show results for eye images with the lowest resolution, that is, with 4×12 pixels.

Estimation Error for Each Positioning Second, we show errors for test images with various positionings in Figure 7 (b). The horizontal axis indicates the index number j for the positionings of the eye regions. Here, $j = 12$ corresponds to the correct positioning. The vertical axis represents the averaged error of five subjects.

Comparing the error at $j = 12$ with those for other indices, it is clear that the PCA-based method is sensitive to positioning errors. On the other hand, the errors of our proposed method are almost the same for all positionings. Therefore, we can conclude that our method is robust against positioning errors.

Estimation Error for Each Individual Finally, we show the estimation error of five subjects A, B, C, D, and E in Figure 8 (a), and the error averaged over the five subjects in (b). This figure shows that the performance of our proposed method is better than that of the PCA-based method for all subjects.

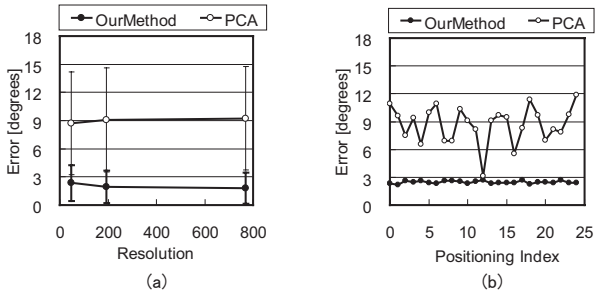

Fig. 7. (a) Errors of gaze estimation against image resolution. (b) Gaze estimation error for each positioning. Index 12 indicates the correct positioning.

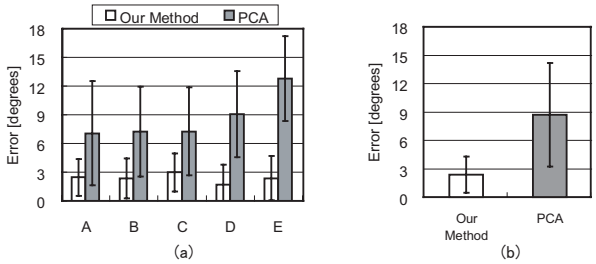

Fig. 8. (a) Gaze estimation errors for each individual and (b) gaze estimation error averaged over all individuals.

Note that the averaged error—2.4 degrees in Figure 8 (b)—is less than half of the sampling distance of the training images—6.4 degrees, the distance between the nearest two circles in Figure 3. The experimental results demonstrate that our bilinear model of two factors, gaze direction and eye region's positioning, can accurately represent the appearance variations resulting from the different gaze directions and positionings.

4 Conclusions

In this study, we proposed a new appearance-based method for gaze estimation from low resolution images, and demonstrated the merit of our proposed method via a number of experiments. One of the key challenges for gaze estimation from low resolution images is that eye regions cannot be found accurately due to limited image resolution, which results in inaccurate estimation of gaze directions. Unlike previously proposed methods, our method is able to estimate gaze directions accurately even when eye regions are found inaccurately in input images.

In order to realize gaze estimation that is insensitive to positioning errors, our method models appearance variation of eye regions due to not only changes in gaze direction but also changes in positioning of eye regions. This is done by incorporating training images of eye regions with artificially added positioning errors, and separating the factor of gaze variation from that of positioning error with a method based on N-mode SVD. In addition, we showed how the problem of gaze estimation can be formulated as a bilinear problem which is solved by alternatively minimizing its cost function with respect to gaze direction and localization of eye regions.

In the present study, we focused on the problem caused by inaccurate positioning of eye regions in low resolution images. Therefore, we did not consider appearance

variations due to other factors such as subject identities and head poses. For our future work, we are planning to extend our method to deal with those factors by incorporating additional modes in the N-mode SVD framework.

References

1. S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. In *Advances in Neural Information Processing Systems*, pages 753–760, 1993.
2. D. Beymer and M. Flickner. Eye gaze tracking using an active stereo head. In *Proc. IEEE CVPR*, pages 451–458, 2003.
3. D. Cristinacce and T. Cootes. Facial feature detection using adaboost with shape constraints. In *Proc. British Machine Vision Conference*, pages 231–240, 2003.
4. T. Hutchinson, K. White JR, W. Martin, K. Reichert, and L. Frey. Human-computer interaction using eye-gaze input. *IEEE Trans. on Systems, Man, and Cybernetics*, 19(6):1527 – 1534, 1989.
5. T. Ishikawa, S. Baker, I. Matthews, and T. Kanade. Passive driver gaze tracking with active appearance models. In *Proc. Intelligent Transportation Systems*, October 2004.
6. Y. Matsumoto and A. Zelinsky. An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In *Proc. IEEE FG*, pages 499–505, 2000.
7. T. Ohno and N. Mukawa. A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. In *Proc. Eye Tracking Research and Application symposium*, pages 115–122, 2004.
8. K. Oka, Y. Sato, Y. Nakanishi, and H. Koike. Head pose estimation system based on particle filtering with adaptive diffusion control. In *IAPR Conf. Machine Vision Applications (MVA 2005)*, pages 586–589, May 2005.
9. H.-Y. Shum, K. Ikeuchi, and R. Reddy. Principal component analysis with missing data and its application to polyhedral object modeling. *IEEE Trans. PAMI*, 17(9):854–867, 1995.
10. R. Stiefelhagen, J. Yang, and A. Waibel. Tracking eyes and monitoring eye gaze. In *Proc. Workshop on Perceptual User Interfaces*, Banff, Canada, October 1997.
11. K.-H. Tan, D. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Proc. IEEE Workshop on Applications of Computer Vision*, pages 191–195, 2002.
12. M. A. O. Vasilescu and D. Terzopoulos. Multilinear image analysis for facial recognition. In *Proc. ICPR*, pages 511–514, 2002.
13. M. A. O. Vasilescu and D. Terzopoulos. Multilinear independent components analysis. In *Proc. IEEE CVPR*, pages 547–553, 2005.
14. J.-G. Wang, E. Sung, and R. Venkateswarlu. Eye gaze estimation from a single image of one eye. In *Proc. IEEE ICCV*, pages 136–143, 2003.
15. L.-Q. Xu, D. Machin, and P. Sheppard. A novel approach to real-time non-intrusive gaze finding. In *Proc. British Machine Vision Conference*, 1998.
16. T. Yao, H. Li, G. Liu, X. Ye, W. Gu, and Y. Jin. A fast and robust face location and feature extraction system. In *Proc. IEEE ICIP*, pages 157–160, 2002.
17. D. Yoo and M. Chung. Non-intrusive eye gaze estimation without knowledge of eye pose. In *Proc. IEEE FG*, pages 785–790, 2004.
18. Z. Zhu and Q. Ji. Eye gaze tracking under natural head movements. In *Proc. IEEE CVPR*, pages 918–923, 2005.