Inferring Human Gaze from Appearance via Adaptive Linear Regression

Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato
Institute of Industrial Science, the University of Tokyo, Japan
{lufeng, sugano, takahiro, ysato}@iis.u-tokyo.ac.jp

Abstract

The problem of estimating human gaze from eye appearance is regarded as mapping high-dimensional features to low-dimensional target space. Conventional methods require densely obtained training samples on the eye appearance manifold, which results in a tedious calibration stage. In this paper, we introduce an adaptive linear regression (ALR) method for accurate mapping via sparsely collected training samples. The key idea is to adaptively find the subset of training samples where the test sample is most linearly representable. We solve the problem via l^1-optimization and thoroughly study the key issues to seek for the best solution for regression. The proposed gaze estimation approach based on ALR is naturally sparse and low-dimensional, giving the ability to infer human gaze from variant resolution eye images using much fewer training samples than existing methods. Especially, the optimization procedure in ALR is extended to solve the subpixel alignment problem simultaneously for low resolution test eye images. Performance of the proposed method is evaluated by extensive experiments against various factors such as number of training samples, feature dimensionality and eye image resolution to verify its effectiveness.

1. Introduction

The goal of gaze estimation technology is to detect human visual attention. Such ability is useful or even crucial in many applications. For instance, it has long been suggested that gaze sensors could be a good alternative to traditional input devices for computers, while commercial systems have already been used for people with disabilities to accomplish tasks only by gazing. Other possible applications are found in various areas including human computer interaction, marketing research, virtual training, and medical research.

Early gaze sensors are intrusive, and thus their usage is severely limited to specific areas. Non-intrusive gaze estimation became a research focus with the development of computer vision technology, where the feature-based methods were first introduced. The most commonly extracted features include corneal infrared reflections, pupil center [13], and iris contour [15]. Using these features, gaze direction is estimated by a wide variety of methods from conventional eyeball modeling [5] to a cross-ratios based geometric method [19]. Comprehensive surveys can be found in [8, 6].

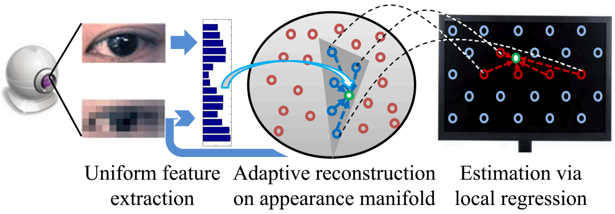

Uniform feature extraction
Adaptive reconstruction on appearance manifold
Estimation via local regression

Figure 1: Overview of the method. Low-dimensional feature is first extracted uniformly from eye images regardless of their resolutions, then reconstructed by training samples adaptively chosen on the appearance manifold via l^1-optimization. The optimization also provides feedback to help extract features from low resolution images with accurate subpixel alignment. Finally, local linear regression is performed to estimate gaze position on the screen.

Unlike feature-based methods, appearance-based methods use an entire eye image as a high-dimensional input feature and map the feature to low-dimensional gaze position space. Their advantage lies in that, as long as there is no need to extract small scale eye features, it would be sufficient to use only a single webcam with relative low resolution instead of using infrared/zoom-in cameras and pan-tilt unit. Baluja and Pomerleau [2] proposed a neural network-based regression method and collected 2000 samples for training. A similar method was also introduced by Xu *et al.* [18]. Tan *et al.* [11] investigated the local linearity on the appearance feature manifold and interpolated unknown gaze position using 252 training samples. However, a major disadvantage of these methods is that obtaining such great number of training samples results in a tedious calibration procedure that cannot be accepted by ordinary users.

Recently, Williams *et al.* [16] introduced a Gaussian Process regression based semi-supervised method to reduce the

978-1-4577-1102-2/11/$26.00 ©2011 IEEE

number of labeled training samples. However, many unlabeled samples are still required. Sugano *et al.* [9] suggested utilizing automatically computed saliency maps to generate training samples while the user is watching a video clip. This method is quite novel but with relatively low accuracy. Some other methods proposed by Sugano *et al.* [10] and Lu *et al.* [7] focus on the specific scenario of free head motion and do not actually reduce the number of training samples for any given head pose. In addition, none of these methods explores the ability of appearance-based methods to use low resolution eye images for gaze estimation.

In this paper, we propose a novel appearance-based method for gaze estimation with sparsely collected training samples. An overview of the method is briefly illustrated in Figure 1. The proposed framework integrates the procedures of feature extraction, adaptive reconstruction on manifold and estimation by regression. Specifically, our method shows the following significant differences from existing methods:

1. To reduce the cost of the cumbersome calibration procedure, the proposed method requires only a small number of sparsely collected training samples, while the estimation accuracy is guaranteed by utilizing adaptive local regression with optimally chosen training samples.

2. The proposed method extracts low-dimensional features uniformly from different resolution eye images. Thus the estimation is resolution-independent. Specifically, accurate estimation is performed with a subpixel alignment method for low resolution test images.

The rest of the paper is organized as follows. Section 2 describes the proposed method in detail. Section 3 presents and discusses the evaluation results from extensive experiments. Finally, Section 4 concludes the paper.

2. Gaze estimation from eye appearance

We propose an appearance-based gaze estimation method aiming at 1) reducing the number of training samples and 2) uniformly supporting different resolution eye images. To achieve these goals, techniques including low-dimensional feature extraction, adaptive linear regression, and subpixel alignment for low resolution test images are introduced under a uniform framework.

2.1. Low-dimensional feature extraction

Existing appearance-based methods generate feature $\boldsymbol{e}_i \in \mathbb{R}^m$ from i-th captured image I_i by raster scan of all pixels, thus the typical feature dimensionality m reaches several thousand [11, 18] or even higher (*e.g.* edge map is added in [16]). High-dimensional feature keeps all the information in the image. However, as eyeball movement is approximately 2-D, such high dimensionality is somewhat redundant. Moreover, the actually captured eye regions should be of variant resolutions, thus pixel-wise feature extraction becomes inapplicable in practice.

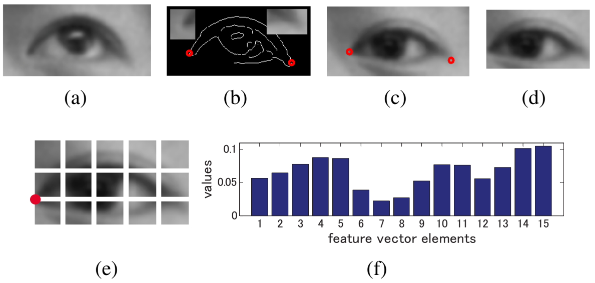


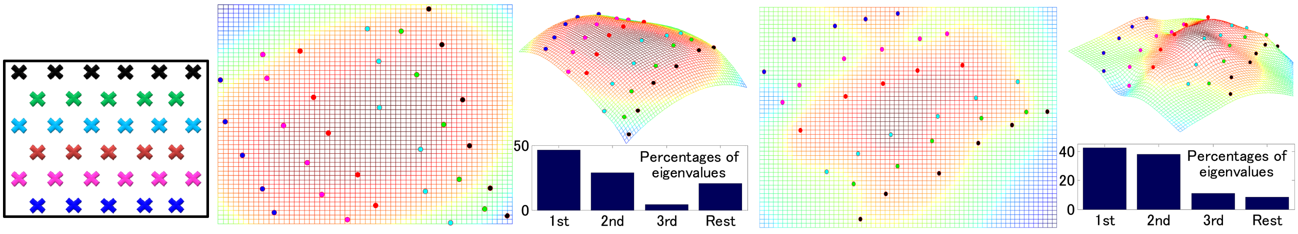
Figure 2: Eye appearance feature extraction. (a) shows the 1st image; (b) shows the detected eye corner and extracted templates; (c)(d) illustrate the eye corner matching and the final crop of the following images; (e)(f) show subregions divided by 3 × 5 and the generated 15-D feature vector.

We introduce a feature extraction method consisting of two steps: eye region crop and feature generation. In the first step, as we assume a fixed head pose, the rough eye regions can be cropped from the original images uniformly, and then bilateral filter is applied (Figure 2a). To deal with small head motion, an additional registration process is performed. For the first eye image, both its inner and outer eye corners are detected by canny edge filter and their corresponding image regions are saved as templates (Figure 2b). For the following eye images, the eye corners are found by matching with the templates (Figure 2c). Finally, the precise eye image I_i is cropped in accordance with the positions of eye corners and a fixed aspect ratio (Figure 2d).

In the feature generation step, the cropped eye image I_i is further divided into $r \times c$ subregions. The c columns are divided evenly, while the r rows are split first from inner eye corners (Figure 2e). Let S_j denote the sum of pixel intensities in j-th subregion, then feature vector is generated by $\boldsymbol{e}_i = \frac{[S_1,S_2,\cdots,S_{r\times c}]^{\mathrm{T}}}{\sum_j S_j}$ (Figure 2f).

2.2. Eye appearance manifold

All the eye appearance features $\{\boldsymbol{e}_i\} \in \mathbb{R}^m$ constitute a manifold in the m-D space. As the eyeball movement has only 2 degrees of freedom, the manifold has an approximately 2-D surface. To test this statement, we project the manifold into 3-D space by PCA transformation for visualization, as shown in Figure 3. Several observations can be obtained: 1) the manifold can be approximated as a 2-D surface with most information accumulating inside two major dimensions; 2) the proposed 15-D feature keeps more information in the first three major dimensions (Figure 3c) than the pixel-wise extracted feature (Figure 3b). Note that



(a) (b) (c)

Figure 3: 2-D gaze space and eye appearance feature manifold. (a) plots the positions of 33 training samples on the 2-D screen plane. (b) shows the eye appearance manifold consisting of pixel-wise extracted features (circles) in 3-D space. The eigenvalues for the three displayed dimensions are also shown in percentage. Similarly, (c) illustrates the proposed 15-D feature case. Notice the similarity of samples' locations between gaze position plane and the manifold surface.

although the manifold in Figure 3b seems smoother, nonlinearity can be hidden in other dimensions; and 3) most importantly, note that after being projected, samples' relative locations on manifold (Figure 3b and 3c) are similar to the gaze positions (Figure 3a).

These observations help us understand the efficiency of linear interpolation-based methods [10, 11]. The basic idea is to reconstruct the appearance feature $\hat{\boldsymbol{e}}$ of the test sample by linear combination. Let $\{\boldsymbol{e}_i\}$ denote the features of training samples and $\{\boldsymbol{x}_i\}$ denote the corresponding gaze positions. Solving

$$\hat{\boldsymbol{e}} = \sum_i w_i \boldsymbol{e}_i \quad s.t. \quad \sum_i w_i = 1 \tag{1}$$

for the weights $\{w_i\}$, then calculate the gaze position $\hat{\boldsymbol{x}}$ by:

$$\hat{\boldsymbol{x}} = \sum_i w_i \boldsymbol{x}_i \tag{2}$$

Let $\{\boldsymbol{e}'_j\}$ and $\{\boldsymbol{x}'_j\}$ denote the subsets of $\{\boldsymbol{e}_i\}$ and $\{\boldsymbol{x}_i\}$ whose corresponding weights $\{w_j \neq 0\}$. Note that using the same $\{w_i\}$ in Equation 1 and 2 requires the linear combination be kept between the subspaces spanned by $\{\boldsymbol{e}'_j\}$ and $\{\boldsymbol{x}'_j\}$. Without any prior knowledge, this requirement is ensured by locality, *i.e.* limiting $\{\boldsymbol{e}'_j\}$ within a sufficient small region. Existing methods [10, 11] determine this local region by selecting $\{\boldsymbol{e}'_j\}$ with the smallest Euclidean distances from $\hat{\boldsymbol{e}}$. However, if the training samples are sparsely collected, as shown in Figure 3a, it is likely that even the nearest sample is not really local. We argue that in such cases, linear methods such as Equation 2 can be still effective. While the key lies on the optimal choice of supporting training samples $\{\boldsymbol{e}'_j\}$ that best keep the linear combination with $\{\boldsymbol{x}'_j\}$.

2.3. Adaptive linear regression via l^1-optimization

Let matrixes $E = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n] \in \mathbb{R}^{m \times n}$ and $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{2 \times n}$ consist of all the features and gaze positions of the training samples. Linear regression can be obtained by simply finding a transformation:

$$AE = X \tag{3}$$

where $A \in \mathbb{R}^{2 \times m}$ is the projection matrix. With training samples' number $n > m$, Equation 3 is overdetermined and each estimation of $\boldsymbol{x}_i$ contains errors. Thus, accurate estimation should be found only for a subset $\{(\boldsymbol{e}'_j, \boldsymbol{x}'_j)\}$. Then local transformation A' is calculated by:

$$A'E' = X' \tag{4}$$

where $E' = [\boldsymbol{e}'_1, \boldsymbol{e}'_2, \cdots, \boldsymbol{e}'_{n'}] \in \mathbb{R}^{m \times n'}$ and $n' < m$.

However, estimation using A' is only accurate for samples included in $\{(\boldsymbol{e}'_j, \boldsymbol{x}'_j)\}$. For any test sample $(\hat{\boldsymbol{e}}, \hat{\boldsymbol{x}})$, there is no guarantee that $A'\hat{\boldsymbol{e}} = \hat{\boldsymbol{x}}$. We consider how to choose $\{(\boldsymbol{e}'_j, \boldsymbol{x}'_j)\}$ in order to make A' most probably accurate for estimating $\hat{\boldsymbol{x}}$ from $\hat{\boldsymbol{e}}$. It is intuitive that $\hat{\boldsymbol{e}}$ should be closely correlated to $\{\boldsymbol{e}'_j\}$ and thus share the same linear mapping. We seek such a 'correlation' by finding the fewest $\boldsymbol{e}'_j$ able to interpolate $\hat{\boldsymbol{e}}$ linearly with total weights equals to 1. Thus $\hat{\boldsymbol{e}}$ is considered to live in the low-dimensional subspace spanned by $\{\boldsymbol{e}'_j\}$ and most probably share the same transformation A' in estimation. Therefore, we calculate:

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_0 \quad s.t. \quad E\boldsymbol{w} = \hat{\boldsymbol{e}} \in \mathbb{R}^m, \mathbf{1}^{\mathrm{T}}\boldsymbol{w} = 1 \tag{5}$$

and then $\{\boldsymbol{e}'_j\}$ is selected from $\{\boldsymbol{e}_i\}$ with nonzero weights in $\hat{\boldsymbol{w}}$. Finally, the gaze position is estimated from $\hat{\boldsymbol{e}}$ by:

$$\begin{aligned} \hat{\boldsymbol{x}} &= A'\hat{\boldsymbol{e}} = A'E\hat{\boldsymbol{w}} = A'E'\hat{\boldsymbol{w}}' + A'E^0\hat{\boldsymbol{w}}^0 \\ &= X'\hat{\boldsymbol{w}}' + \mathbf{0} = X'\hat{\boldsymbol{w}}' + X^0\hat{\boldsymbol{w}}^0 = X\hat{\boldsymbol{w}} \in \mathbb{R}^2 \end{aligned} \tag{6}$$

where $\hat{\boldsymbol{w}}'/\hat{\boldsymbol{w}}^0$ is composed by the nonzero/zero elements in $\hat{\boldsymbol{w}}$, and E^0/X^0 is formed by columns from E/X that correspond to $\hat{\boldsymbol{w}}^0$.

Having Equation 6, we only need to solve Equation 5 for $\hat{\boldsymbol{w}}$ rather than obtaining A'. However, solving this problem is NP-hard [1]. Fortunately, l^1-norm minimization can be an alternative when $\hat{\boldsymbol{w}}$ is sparse enough [3, 4]. At the same time, note that according to Section 2.1, any extracted feature $\boldsymbol{e}_i$ satisfies $\mathbf{1}^{\mathrm{T}}\boldsymbol{e}_i = 1$. Thus from $E\boldsymbol{w} = \hat{\boldsymbol{e}}$, we have

$$\mathbf{1}^{\mathrm{T}}(E\boldsymbol{w}) = \mathbf{1}^{\mathrm{T}}(\hat{\boldsymbol{e}}) \Rightarrow [\cdots, \mathbf{1}^{\mathrm{T}}\boldsymbol{e}_i, \cdots]\boldsymbol{w} = 1 \Rightarrow \mathbf{1}^{\mathrm{T}}\boldsymbol{w} = 1 \tag{7}$$

which naturally ensures the second constraint in Equation 5 and this constraint can be omitted. In addition, a tolerance term ε should be introduced to make a tradeoff between sparsity and linear combination precision:

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_1 \quad s.t. \quad \|E\boldsymbol{w} - \hat{\boldsymbol{e}}\|_2 < \varepsilon \tag{8}$$

By minimizing $\|\boldsymbol{w}\|_1$, we are actually pursuing the lowest dimensional subspace where the interpolation result $E\boldsymbol{w}$ lives. Note that ε restricts the maximum allowed Euclidean distance from $E\boldsymbol{w}$ to the true $\hat{\boldsymbol{e}}$. Thus with small ε, $\hat{\boldsymbol{e}}$ is considered to be in the same subspace approximately, and local transformation in Equation 4 is applicable. As an important complement of the proposed method, the optimal choice of ε is thoroughly investigated and given later in Section 3.1 along with experimental demonstrations.

Relationship to other works. l^1-minimization has been used in many vision-related researches, among which the works by Wright *et al.* [17] and Tan *et al.* [12] share similarities with ours. However, our work essentially differs in two main ways. First, as both their problems concern face recognition, they are more focused on discriminability; while our purpose is to find the best linear regression for mapping from feature space to target space (*i.e.* the 2-D gaze position space). Second, to seek sparsity, Wright *et al.* [17] also introduced an error term ε while not mentioning how to determine its value. However, as discussed later in Section 3.1, the value of ε is essential and meaningful in our case and must be carefully chosen. Furthermore, Tan *et al.* [12] transformed the data using a Gaussian random matrix in order to meet the RIP condition [3] for sparsity. This is not applicable in our situation because the transformation is not affine. Thus linear combination weights cannot be kept between transformed feature space and original target space and also Equation 7 fails.

2.4. Subpixel alignment for low resolution image

As shown in Figure 4, for feature extraction, the method in Section 2.1 positions the extracting region by detecting two eye corners. However, with a low resolution eye image, it is difficult to find eye corners with sufficient precision. Here we propose an alignment method to accurately extract features from low resolution eye regions.

Assuming frontal face orientation, we first estimate the extracted region size via large scale face features. Then what must be determined is the alignment position (x, y). Let $\boldsymbol{e}^{(x,y)}$ denote the feature extracted with the alignment position (x, y). If all training samples $\{\boldsymbol{e}_i\}$ are generated from accurately cropped eye images, then only when (x, y) is also correctly found, $\boldsymbol{e}^{(x,y)}$ lies on the appearance manifold constituted by $\{\boldsymbol{e}_i\}$ and $\|\boldsymbol{w}\|_1$ achieves the minimum of 1. Figure 5a shows an example where $\|\boldsymbol{w}\|_1$ becomes the global minimum with accurate alignment position (x, y). Thus we optimize (x, y) by minimizing $\|\boldsymbol{w}\|_1$:

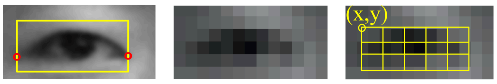
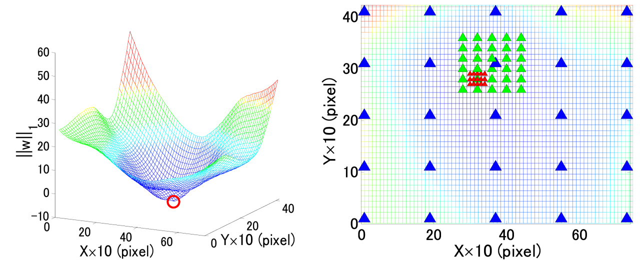


Figure 4: Eye image alignment for feature extraction. The feature extraction method in Section 2.1 is based on the precise eye corner detection, which is not feasible for low resolution images. In that case, the alignment position (x, y) should be determined with subpixel accuracy.



(a) (b)

Figure 5: Subpixel alignment by minimizing $\|\boldsymbol{w}\|_1$. (a) shows accurate alignment position with minimized $\|\boldsymbol{w}\|_1$; (b) illustrates the hierarchical search procedure of 3 levels for the minimized $\|\boldsymbol{w}\|_1$. Markers with different sizes/colors indicate searching positions in each level.

$$((\hat{x}, \hat{y}), \hat{\boldsymbol{w}}) = \arg\min_{(x,y),\boldsymbol{w}} \|\boldsymbol{w}\|_1 \; s.t. \; \|E\boldsymbol{w} - \boldsymbol{e}^{(x,y)}\|_2 < \varepsilon \tag{9}$$

To efficiently find the solution of Equation 9, hierarchical search is used to solve the problem with required subpixel accuracy. First, candidates for (x, y) are sparsely chosen among all the feasible alignment positions. After l^1-minimization, only the candidate with smallest $\|\boldsymbol{w}\|_1$ is selected as the start point for next search level. For each level, the search range decreases while the precision increases. A threshold for the minimum search step is preset to terminate the procedure when desired precision is reached. An example of the search procedure is demonstrated in Figure 5b, where alignment accuracy of 0.1 pixel is achieved via 3 levels search. Note that if the total number of feasible positions at desired precision level is N, then computational complexity is reduced to $M \log_M N$ approximately, where M is the number of candidates chosen in each level.

The solution of Equation 9 is usually accurate enough for estimation. If higher accuracy is required, the obtained

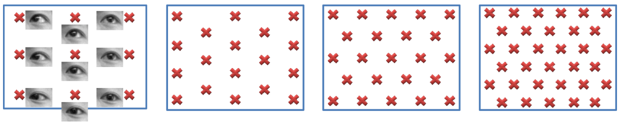


Figure 6: Patterns for obtaining training samples. Red crosses indicate the calibration points on the screen with total number of 9, 18, 23, and 33 in each pattern.

$(\hat{x}, \hat{y})$ and the corresponding $(x^{\lrcorner}, y^{\lrcorner})$, which denotes the bottom right position of the alignment region, can be set as initial values. Then further alignment can be done by using a method such as gradient descent to optimize these two positions with a cost function similar to Equation 9.

Wagner *et al.* [14] also proposed to use l^1-minimization for alignment with regard to the face recognition problem. However, their method only involves face images from the same person to avoid local minimum. Thus it minimizes the l^1-norm of the reconstruction error rather than the weights because the weights are dense for the same person. On the contrary, our method directly minimizes the l^1-norm of the weights for the global minimum. Furthermore, their method is designed for pixel-wise alignment of high resolution face image, while in our case we align with subpixel accuracy for low resolution eye image.

3. Evaluation and discussion

We evaluated the performance of the proposed method via extensive experiments. To obtain training and test data, we developed a system on a desktop PC with an 18-inch LCD monitor and a webcam of VGA resolution. The user was asked to sit in front of the monitor (about 50cm-60cm away) and keep his head stable with the help of a chinrest. The system captured the user's frontal appearance while his gaze was focusing on each marker shown on the screen. The markers' positions were saved and the appearance features were extracted from the captured images.

As one of our goals is to allow sparse sampling of training data, we designed several sampling patterns on the screen that contained only 9, 18, 23, or 33 training points, as shown in Figure 6.

To measure the accuracy, estimation error is calculated by view angle:

$$error = \arctan(\frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}'\|_2}{d_{user}}) \tag{10}$$

where $\|\hat{\boldsymbol{x}} - \boldsymbol{x}'\|_2$ denotes the Euclidean distance between real gaze position $\boldsymbol{x}'$ and the estimated gaze position $\hat{\boldsymbol{x}}$, and d_{user} indicates the average distance from user's eyes to the screen. In practice, gaze position is computed as the average estimation results for both left and right eyes.

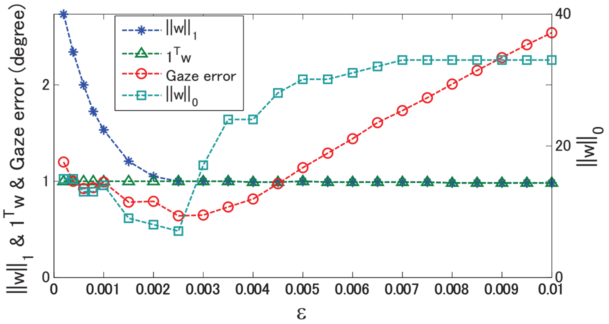


Figure 7: An example of solved weights $\hat{\boldsymbol{w}}$ and gaze estimation error under different ε for one test sample.

3.1. Determining ε for l^1-optimization

We complete the discussion on determining the value for ε in Section 2.3. Intuitively, larger ε results in smaller $\|\hat{\boldsymbol{w}}\|_1$ from Equation 8. Then essential issues include how is the proper value of ε determined? What is more important, does this process keep reducing $\|\hat{\boldsymbol{w}}\|_0$? We show one typical experimental result in Figure 7. It is clear that with ε increasing, $\|\hat{\boldsymbol{w}}\|_0$ reaches the minimum when $\|\hat{\boldsymbol{w}}\|_1$ just converges to $\mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{w}} = 1$. At the same time, the estimation error is also minimized. If ε continues to enlarge, both $\|\hat{\boldsymbol{w}}\|_0$ and the estimation error increase.

To understand this, note that if Equation 7 holds, then

$$\|\boldsymbol{w}\|_1 \geq \mathbf{1}^{\mathrm{T}}\boldsymbol{w} = 1 \tag{11}$$

meaning that $\|\hat{\boldsymbol{w}}\|_1$ reaches the minimum of 1 when all the elements in $\hat{\boldsymbol{w}}$ become nonnegative. After this, $\|\hat{\boldsymbol{w}}\|_1$ equals $\mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{w}}$, thus further increasing ε turns Equation 8 into

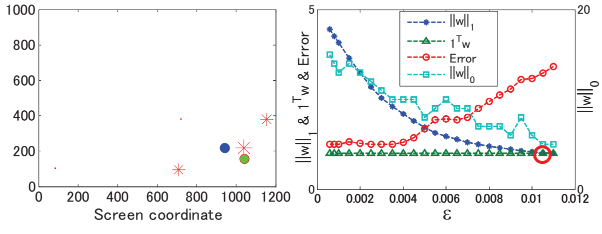
$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \mathbf{1}^{\mathrm{T}}\boldsymbol{w} \quad s.t. \quad \|E\boldsymbol{w} - \hat{\boldsymbol{e}}\|_2 < \varepsilon \tag{12}$$

According to Equation 7, $\mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{w}}$ is directly controlled by $E\boldsymbol{w} - \hat{\boldsymbol{e}}$[1]. Therefore, Equation 12 no longer reduces $\|\hat{\boldsymbol{w}}\|_0$. Instead, $\hat{\boldsymbol{w}}$ becomes dense and $\|\hat{\boldsymbol{w}}\|_0$ increases. Moreover, Equation 12 breaks the constraint of $\mathbf{1}^{\mathrm{T}}\boldsymbol{w} = 1$ and makes the linear combination invalid for estimation.

Therefore, the optimal value of ε should be determined when the minimized $\|\hat{\boldsymbol{w}}\|_1$ just converges to 1, which brings two benefits: 1) such ε achieves the sparsest solution $\hat{\boldsymbol{w}}$ which is obtained by Equation 8 rather than Equation 12, while not breaking the constraint of $\mathbf{1}^{\mathrm{T}}\boldsymbol{w} = 1$; and 2) the nonnegativity of $\hat{\boldsymbol{w}}$ ensures that the linear combination is performed inside the convex region constituted by $\{\boldsymbol{e}'_j\}$, which is a good property for interpolation.

However, optimizing ε for every test sample is time consuming. On the basis of observation that the optimal value

[1] $\mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{w}} = \mathbf{1}^{\mathrm{T}}(E\boldsymbol{w} - \hat{\boldsymbol{e}}) + 1$



(a) (b)

Figure 8: Example of estimating peripheral gaze position. (a): the true gaze position (green circle) lies out of the region composed by the supporting training samples (red stars), and the interpolation result (blue circle) with all-nonnegative weights is not accurate. Thus an extrapolation should be better with negative weights. (b): therefore, if we still choose a value for ε when $\|\hat{\boldsymbol{w}}\|_1$ converges to 1 (indicated by the circle), it will be too large for estimation.

of ε is very stable under the same parameters, it is better to use a constant value ε_c for all the test samples:

$$\varepsilon_c = \frac{\sum_i \alpha_i \varepsilon_i}{\sum_i \alpha_i}, \quad \alpha_i = \exp(-\kappa \|\boldsymbol{x}_i - \dot{\boldsymbol{x}}\|_2) \tag{13}$$

where ε_i is obtained by leave-one-out experiments for each training sample, and $\dot{\boldsymbol{x}}$ is the screen center position. With sufficiently large κ, the calculated weights $\{\alpha_i\}$ in Equation 13 ensure that ε_i from the peripheral sample is less influential to ε_c. This is illustrated and explained in Figure 8.

3.2. Comprehensive evaluation and comparison

We first evaluated the estimation accuracy using our datasets. Training samples were obtained as shown in Figure 6 with total numbers of 9, 18, 23, and 33. Test samples were evenly collected on the screen without overlapping the training samples. Eye appearance features were extracted using the method in Section 2.1 with 2 × 4 and 3 × 5 subregion division strategies, denoted as 8-D and 15-D feature, respectively. A total of 332 training samples and 800 test samples were tested for four subjects.

Table 1 compares the estimation errors of the proposed method with those of the commonly used local region-based method [11]. The local region-based method finds the closest triangle to the test sample in the Delaunay triangulated mesh, and chooses its three vertex and their direct neighbors as supporting training samples. We also modified the local region-based method by only selecting the three vertex of the closest triangle to be the supporting samples (denoted as 'local triangle region-based'). This modification improved the performance because the training samples were sparsely collected in our datasets. The results demonstrate that the proposed method achieves the highest estimation accuracy in most experimental conditions. Note that solving Equation 8 requires smaller feature dimensionality than the training samples' number, thus the results for '15-D/9' are not obtained. Figure 9 plots the average estimation errors and standard deviations to show the advantage of the proposed method intuitively.

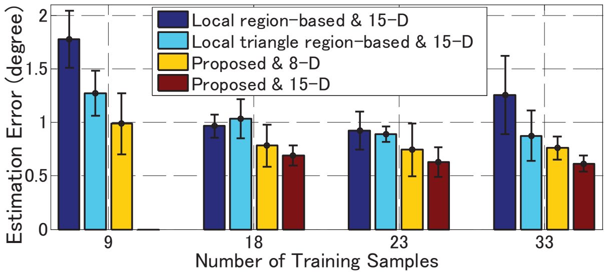


Figure 9: Comparison of average gaze estimation errors and standard deviations from Table 1.

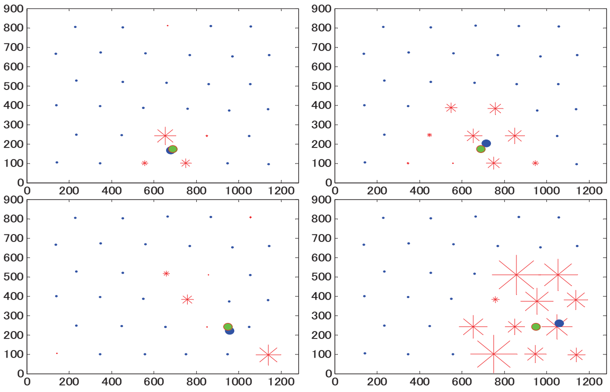


Figure 10: Two estimation examples are illustrated in each row. Left: estimated gaze positions (blue circles) by the proposed method and the true gaze positions (green circles) in screen coordinate; Right: comparative results by local region-based method [11]. Specifically, the red stars indicate the selected training samples with their sizes indicating the absolute values of linear combination weights.

Figure 10 shows two individual estimation examples to further explain the benefit of the proposed method. The local region-based method clearly assigns weights densely to the supporting samples, while the proposed method adaptively selects the fewest samples through all training samples in accordance with linear representability rather than distance, and achieves better estimation accuracy.

Finally, as additional information, Table 2 compares the accuracy of our method with those reported by existing appearance-based gaze tracking approaches. Note that this comparison is not direct due to the fact that different systems require substantially different training data. However, it can still be revealed from Table 2 that our method shows the ability to accept much fewer training samples while

Feature dimensionality/training samples		8-D/9	8-D/18	8-D/23	8-D/33	15-D/9	15-D/18	15-D/23	15-D/33
Subject A	Proposed method	**1.03**	**1.07**	**0.91**	**0.85**	-	**0.75**	**0.70**	**0.66**
	Local region-based [11]	3.05	2.02	1.66	1.72	1.91	1.11	0.94	1.11
	Local triangle region-based	1.61	1.15	1.27	1.52	**0.97**	0.82	0.83	1.16
Subject B	Proposed method	**0.99**	**0.80**	1.03	**0.88**	-	**0.80**	**0.76**	**0.67**
	Local region-based [11]	5.11	2.76	1.92	2.15	**1.35**	0.81	1.20	1.12
	Local triangle region-based	1.22	1.02	**0.93**	1.22	1.40	1.15	0.84	0.74
Subject C	Proposed method	**1.34**	**0.74**	**0.63**	**0.68**	-	**0.64**	**0.66**	**0.63**
	Local region-based [11]	2.52	1.72	2.33	2.00	1.84	0.94	0.75	1.32
	Local triangle region-based	1.47	1.16	0.79	1.19	**1.20**	0.95	0.98	0.80
Subject D	Proposed method	**0.59**	**0.52**	**0.40**	**0.63**	-	**0.57**	**0.40**	**0.50**
	Local region-based [11]	3.23	1.85	1.89	1.15	2.00	1.00	0.80	1.47
	Local triangle region-based	1.02	0.55	0.47	0.63	**1.52**	1.22	0.90	0.80
Average	Proposed method	**0.99**	**0.78**	**0.74**	**0.76**	-	**0.69**	**0.63**	**0.62**
	Local region-based [11]	3.48	2.09	1.95	1.76	1.78	0.97	0.92	1.26
	Local triangle region-based	1.33	0.97	0.87	1.14	**1.27**	1.04	0.89	0.88

Table 1: Comparison of gaze estimation error (in degree) between the proposed method, the original and modified local region-based methods using the same datasets. Totally 332 training samples and 800 test samples are tested for 4 subjects.

Method	Error	Training samples
Proposed	0.99°	9
Proposed	0.69°	18
Proposed	0.62°	33
S^{3}GP [16]	1.32°	16 labeled and 75 unlabeled
S^{3}GP+edge+filter [16]	0.83°	16 labeled and 75 unlabeled
Tan *et al.* [11]	0.5°	252
Baluja *et al.* [2]	1.5°	2000
Xu *et al.* [18]	1.5°	3000

Table 2: Comparison with existing appearance-based methods based on their reported estimation error and number of training samples .

achieving very high accuracy. Using 15-D feature and 18 training samples is a good tradeoff between easy calibration and high precision. Note that the method of Tan *et al.* [11] achieves the highest accuracy in Table 2 with 252 training samples. However, given much fewer training samples with low dimensionality, its accuracy degrades significantly in our experiments as shown in Table 1 and Figure 9. Other notable points include 1) all these existing methods use very high dimensional features generated from high resolution eye images, while our method only extracts very low dimensional feature regardless of eye image resolution; and 2) our method adaptively selects supporting training samples, thus is the only one that allows directly adding/removing any training samples into/from the system.

3.3. Performance for low resolution images

Performance of using high resolution training images to estimate gaze from low resolution test images is assessed.

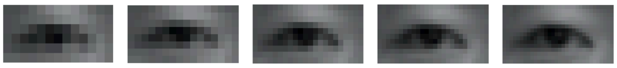
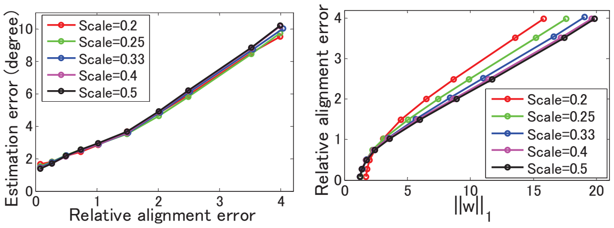


Figure 11: Low resolution images by down-sampling. The vertical and horizontal scaling factor for each image is 0.2, 0.25, 0.33, 0.4 and 0.5, respectively.

The low resolution images can be acquired either by low resolution cameras or capturing from a distance. However, to obtain the ground truth, down-sampled images were used in our experiment by scaling factors of 0.2, 0.25, 0.33, 0.4, and 0.5 for both vertical and horizontal directions. Examples are illustrated in Figure 11, where each eye region includes around 30, 50, 90, 110, and 170 pixels. Features were extracted by the method introduced in Section 2.4.

Experiments were conducted under different scaling factors. Figure 12a plots how gaze estimation error varies with relative alignment error (misalignment in pixels/scaling factor). It is clear that smaller alignment error reduces estimation error directly (the lower bound is less than 2°). Besides, as shown in Figure 12b, the alignment error is closely correlated to $\|\hat{\boldsymbol{w}}\|_1$ and tends to be zero when the minimum of $\|\hat{\boldsymbol{w}}\|_1 = 1$ is reached by the optimization introduced in Section 2.4. These experimental evidences indicate that by minimizing $\|\hat{\boldsymbol{w}}\|_1$ our method is able to achieve the best estimation accuracy. Note that these results show little sensitivity to different scaling factors (except 0.2 which is extremely small), which helps to demonstrate the effectiveness of the proposed alignment method and low-dimensional feature. The final results for gaze estimation from low resolution test images are shown in Figure 13.



(a) (b)

Figure 12: Relationships between alignment error, estimation error, and $\|\hat{w}\|_1$ minimization.

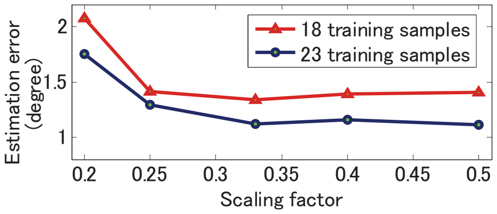


Figure 13: Results of gaze estimation using low resolution images. Note that the accuracy is still comparable to other high resolution and dense sampling methods.

3.4. Discussion of limitation

Our method accepts low resolution images as test inputs, which in practice are supposed to be captured from a distance. On the other hand, like most existing appearance-based methods, our method currently assumes a fixed head pose near the screen. Thus there is still way to go towards practical gaze tracking systems by combining the proposed resolution-invariant technique with head pose-free scheme, which is our future research direction.

4. Conclusion

We propose a novel approach for human gaze estimation from eye appearance. The approach is built upon the adaptive linear regression method that automatically selects training samples for mapping. Compared with commonly used local region-based regression, our method achieves higher accuracy while using fewer training samples. Furthermore, the proposed low-dimensional feature extraction and subpixel alignment techniques enable accurate estimation even for very low resolution images. On the basis of the proposed method, a gaze tracking system can be implemented that allows easy calibration and is robust to both resolution variation and misalignment.

References

[1] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1-2):237–260, 1998. 4

[2] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. In *NIPS*, volume 6, 1994. 1, 7

[3] E. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory,*, 52(12):5406–5425, 2006. 4

[4] D. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006. 4

[5] E. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans. on Biomedical Engineering*, 53(6):1124–1133, 2006. 1

[6] D. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *PAMI*, 32(3):478–500, 2010. 1

[7] F. Lu, T. Okabe, Y. Sugano, and Y. Sato. A head pose-free approach for appearance-based gaze estimation. In *BMVC*, 2011. 2

[8] C. Morimoto and M. Mimica. Eye gaze tracking techniques for interactive applications. *CVIU*, 98(1):4–24, 2005. 1

[9] Y. Sugano, Y. Matsushita, and Y. Sato. Calibration-free gaze sensing using saliency maps. In *CVPR*, pages 2667–2674, 2010. 2

[10] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike. An incremental learning method for unconstrained gaze estimation. In *ECCV*, pages 656–667, 2008. 2, 3

[11] K. Tan, D. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *WACV*, pages 191–195, 2002. 1, 2, 3, 6, 7

[12] X. Tan, L. Qiao, W. Gao, and J. Liu. Robust faces manifold modeling: Most expressive Vs. most Sparse criterion. In *ICCV Workshops*, pages 139–146, 2010. 4

[13] R. Valenti and T. Gevers. Accurate eye center location and tracking using isophote curvature. In *CVPR*, pages 1–8, 2008. 1

[14] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma. Towards a practical face recognition system: robust registration and illumination by sparse representation. In *CVPR 2009*, pages 597–604, 2009. 5

[15] J. Wang, E. Sung, and R. Venkateswarlu. Eye gaze estimation from a single image of one eye. In *ICCV*, pages 136–143, 2003. 1

[16] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the S^{3}GP. In *CVPR*, pages 230–237, 2006. 1, 2, 7

[17] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 31(2):210–227, 2008. 4

[18] L. Q. Xu, D. Machin, and P. Sheppard. A novel approach to real-time non-intrusive gaze finding. In *BMVC*, pages 428–437, 1998. 1, 2, 7

[19] D. Yoo and M. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *CVIU*, 98(1):25–51, 2005. 1