This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Learning Reconstruction-based Remote Gaze Estimation

Pei Yu, Jiahuan Zhou and Ying Wu
Northwestern University
2145 Sheridan Road, Evanston IL, 60208
{pyi980, jzt011, yingwu}@eecs.northwestern.edu

Abstract

It is a challenging problem to accurately estimate gazes from low-resolution eye images that do not provide fine and detailed features for eyes. Existing methods attempt to establish the mapping between the visual appearance space to the gaze space. Different from the direct regression approach, the reconstruction-based approach represents appearance and gaze via local linear reconstruction in their own spaces. A common treatment is to use the same local reconstruction in the two spaces, i.e., the reconstruction weights in the appearance space are transferred to the gaze space for gaze reconstruction. However, this questionable treatment is taken for granted but has never been justified, leading to significant errors in gaze estimation. This paper is focused on the study of this fundamental issue. It shows that the distance metric in the appearance space needs to be adjusted, before the same reconstruction can be used. A novel method is proposed to learn the metric, such that the affinity structure of the appearance space under this new metric is as close as possible to the affinity structure of the gaze space under the normal Euclidean metric. Furthermore, the local affinity structure invariance is utilized to further regularize the solution to the reconstruction weights, so as to obtain a more robust and accurate solution. Effectiveness of the proposed method is validated and demonstrated through extensive experiments on different subjects.

1. Introduction

Visual sense is the most informative one among all human perceptions. Gaze estimation, which infers visual attention, has wide applications in many areas, *e.g.*, human-computer interfaces can utilize estimated gaze as an auxiliary input to aid disabled people.

Eye gaze tracking techniques can be divided into two categories, intrusive and non-intrusive eye gaze trackers [12]. Intrusive eye gaze trackers provide high accuracy and reliability, but require dedicated hardware. One example uses contact lens, and the gaze directions are estimated by measuring the change of voltage in the lens.

Non-intrusive eye gaze trackers are mostly vision-based. They utilize the information from remote cameras, thus are easier and suitable to use for a long period of time. These techniques can be feature-based or appearance-based. Feature-based methods track eye features such as iris contour, pupil location and *etc*. One popular method uses infrared light source to create glints on cornea.

Since the feature-based methods still need expensive devices, appearance-based methods, which only need one simple web camera to capture eye appearances, is simpler and more attractive in practice. Such methods attempt to construct the mapping between the visual appearance of the eyes and the gaze. In [17] [9], the same weights that reconstruct the visual appearance are used to reconstruct the gaze. These reconstruction-based methods are used as the base estimation when handling the head movements and simplifying the calibration process [8] [10] [16], and they have shown promising performance and considerable simplicity.

However, there is one fundamental problem in such reconstruction-based methods that has not been well studied yet. The using of the same reconstruction weights in the two different spaces (*i.e*. the appearance space and the gaze space) is taken for granted, without justification or guarantee. In fact, they may not be the same, and can even be significantly different. Without addressing this issue, regardlessly using the same reconstruction weights leads to inaccurate results. This issue has not been investigated in the literature.

To address this issue, we propose a new solution that explicitly decomposes the gaze estimation error of the reconstruction-based methods into two terms. The first term is the matching error between the reconstructed appearance and the query appearance. The second one is the error incurred by using the same reconstruction weight in the appearance space and the gaze space, which has never been investigated before. We reduce the second error term via exploring the relationship between the reconstruction weights in the two spaces by analyzing their structures. Two novel methods, learning a global distance metric in training phase and minimizing the local space structure discrepancy in estimation phase, are proposed to align the two spaces, so as to reduce the second error term. The effectiveness of these two new methods is validated through their consistently and significantly superior performances in our experiments on different subjects. The average estimation error of our proposed method outperforms existing methods by at least 11%. The proposed method also shows resilience to pose and subject variation in cross pose and cross subject experiments.

Our proposed method has following significant differences from existing methods:

1. This is the first work to investigate the difference of the reconstruction weights in appearance space and gaze space, and the estimation error incurred by this difference.
2. Two novel methods are proposed to reduce this error term by aligning the two spaces. The first one is used in training phase as a global distance metric learning. The second one is conducted in the estimation phase by minimizing the discrepancy of local space structure between the query appearance and the estimated gaze.

This paper is organized as follows. Section 2 briefly describes the related works. Section 3 describes our proposed method. Section 4 assesses the proposed method, and the baseline methods, with different settings. Section 5 concludes this paper.

2. Related work

Gaze estimation methods can be categorized from two perspectives. In term of the visual features used, there are feature-based methods that utilize specific eye features, and appearance-based methods that treat eye images as high dimensional features. From the perspective of computing, there are model-based methods that reconstruct the 3-D eye model from the input features, and regression methods that learn a mapping directly from the input feature to the gaze. Comprehensive surveys can be found in [7] [12].

Feature-based method: Feature based methods extract specific eye features such as Pupil Center Corneal Reflection (PCCR) [24] [25] [23] [22] [3] [6], iris contour [18], eye corners [2]. Zhu and Ji obtained a head mapping function to compensate head movements [24]. In [25], support vector regression (SVR) is utilized. In [23] [22], Yoo *et al.* and subsequently Yoo and Chung proposed to use five infra red light sources and cross ratio to allow head motion. In [3], Chen and Ji proposed to utilize saliency map to incrementally learn a distribution of the person dependent parameters and the gaze. A detailed study of the methods using PCCR can be found in [6]. All of above methods have to use infra red camera to capture cornea glint. In [18], Wang *et al*. proposed to extract the iris circle from natural image. In [2], Chen and Ji used facial feature points. Although the feature-based methods may provide accurate estimation of gaze, they require infra red devices or high resolution imagery, which limits their adaptability to outdoor situation and/or low resolution camera.

Appearance-based method: Appearance-based methods simply use eye images as high dimensional input features, which relieves gaze estimation from relying on dedicated hardware. In [1] [21], neural network was applied to learn the regression. To reduce the number of training samples, Tan *et al*. proposed to estimate gaze using Locally Linear Reconstruction (LLR) [17] [13]. This method assumes an appearance manifold, where the k-nearest neighbors (k-NN) of one appearance vector are used to form the Delaunay triangulation topology of the corresponding gaze positions. Such a reconstruction is assumed to be the same in both the appearance manifold and the gaze manifold. Lu *et al*. proposed Adaptive Linear Regression (ALR), leveraging the sparsity to choose supporting training samples [9]. In [20], Williams *et al.* proposed a sparse semi-supervised Gaussian process regression model. To handle the head movements, Sugano *et al.* proposed an incremental learning method [16]. Lu *et al.* proposed a two stage solution, i.e., the initial gaze estimation under a fixed head pose and the compensation of the estimation bias caused by the head movement [8]. Furthermore, Lu *et al.* proposed a synthesis approach to generate training images of unseen head poses [10]. Depth camera was utilized by Funes and Odobez [11]. To simplify the training procedure, Sugano *et al*. proposed to use saliency map [14]. For the issue of cross subject estimation, Sugano *et al.* applied the learning-by-synthesis approach and the redundant regression forests [15].

The proposed method is appearance-based, but focuses on studying the relation and alignment between the appearance space and the gaze space. The unclear relation between the two spaces is a significant factor leading to the estimation error in the reconstruction-based methods, but this issue has not been addressed before.

3. The proposed approach

3.1. Baseline methods

Appearance-based gaze estimation is to infer the gaze position from eye appearances through learning a mapping $f : \alpha \mapsto \beta$, where α is the space of appearance features $\mathbf{a} \in \mathbb{R}^n$, and β is the space of gaze position vectors $\mathbf{y} \in \mathbb{R}^d$. A set of training pairs, or called exemplars, $(\mathbf{a}_i, \mathbf{y}_i), i = 1 \ldots N$ is needed. Without losing generality, we represent a gaze $\mathbf{y}$ by a 2-D position on the target plane.

One solution is to learn a global regression function, *e.g.* Support Vector Regression (SVR), that constructs the direct mapping from $\mathbf{a}$ to $\mathbf{y}$. However, such regression relationship between $\mathbf{a}$ and $\mathbf{y}$ can be very complicated, and very difficult to capture, especially when the number of training exemplars is small.

Unlike the direct regression methods that do not use the structure information of the appearance space α, an alternative approach takes advantage of the local structures. It reconstructs the query appearance feature based on the training appearance exemplars (*e.g.*, using the k-nearest neighbors of the query appearance):

$$\omega_\alpha^* = \arg\min_\omega \| \mathbf{a} - \mathbf{A}\omega \|_2^2 \quad s.t.\ \mathbf{1}^T\omega = 1 \tag{1}$$

where $\mathbf{a}$ is query appearance feature, and $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ is the set of the related exemplars in the training set. They can be simply the nearest neighbors [17], or those that give the most sparse reconstruction [9]. Assuming that such reconstruction holds the same in both the appearance space and the gaze space, this approach uses the same weight ω_α^* to interpolate the gaze by the corresponding gaze exemplars:

$$\mathbf{y} = \mathbf{B}\omega_\alpha^* \tag{2}$$

where $\mathbf{B} = \{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$, and each $\mathbf{y}_i$ is the corresponding gaze of its appearance $\mathbf{a}_i$ in $\mathbf{A}$.

3.2. Decomposing the estimation error

By minimizing the objective function in Equ. 1, those traditional methods seek a best reconstructed appearance in the appearance space to best match the query appearance $\mathbf{a}$, *i.e*. minimizing the matching error $\mathcal{D}(\mathbf{a}, \mathbf{A}\omega)$. In [9] [17], $\mathcal{D}(\mathbf{a}, \mathbf{A}\omega) = \| \mathbf{a} - \mathbf{A}\omega \|_2$. This seems to be plausible due to the correspondences between appearance $\mathbf{a}$ and gaze position $\mathbf{y}$, hoping the true gaze to be recovered via transferring the reconstruction. Unfortunately, such correspondences do not automatically imply or guarantee that the reconstruction weights are the same in the two spaces. Regardlessly using the same reconstruction of the query appearance to reconstruct the estimated gaze will introduce significant estimation error.

In fact, the reconstruction in the gaze space β should be given by:

$$\omega_\beta^* = \arg\min_\omega \| \mathbf{y} - \mathbf{B}\omega \|_2^2 \quad s.t.\ \mathbf{1}^T\omega = 1 \tag{3}$$

where $\mathbf{y}$ is the true gaze of query appearance $\mathbf{a}$. Previous works all took for granted that $\omega_\beta^* = \omega_\alpha^*$, which is in fact not true in practice. As shown in Equ. 1, weight ω_α^* depends on the distance between different appearance feature vectors $\mathcal{D}(\mathbf{a}_i, \mathbf{a}_j)$ that exhibits the structure of the appearance space α. However, as illustrated in Fig. 1, the structure of the appearance space α and the gaze position space β may be very different. Because of this discrepancy between the structure of the two spaces, simply using ω_α^* to reconstruct the estimated gaze will inevitably introduce error in gaze estimation.

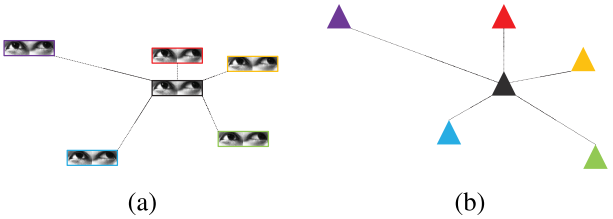


Figure 1. Illustration of the difference of structure between the appearance feature space and the gaze position space. The length of connection lines indicates the Euclidean distance between two connected components. Figure (a) describes the Euclidean distances between the appearance bounded by black box and its 5-nearest neighbors. Figure (b) describes the distances in the gaze position space, where each triangle indicates the gaze position point corresponding to the appearance bounded by the box with the same color.

Therefore, in the reconstruction-based methods, the estimation error comes from two sources, and we explicitly decompose it as:

$$E(\omega) = E_D(\omega) + E_R(\omega) \tag{4}$$

where $E(\omega)$ is the total estimation error using ω to reconstruct estimated gaze, $E_D(\omega)$ is the matching error between the reconstructed appearance using ω and the query appearance, and $E_R(\omega)$ indicates the error introduced by using the appearance reconstruction weight ω to reconstruct the gaze. In previous works of appearance based gaze estimation, only the appearance reconstruction error $E_D(\omega)$ is considered. $E_R(\omega)$ term has not been investigated before, but is critical to bridging the gap of ω between the appearance space α and the gaze space β.

In the following sections, we propose two novel methods to reduce $E_R(\omega)$ from two different perspectives. The first one is to find a global alignment between the appearance space α and the gaze space β via learning a new distance metric in Sec. 3.3. The second one is locally regularizing the discrepancy of space structure related to query appearance $\mathbf{a}$ and estimated gaze $\mathbf{B}\omega$ in Sec. 3.4. The first one is conducted in training phase, while the second one is conducted in estimation phase.

3.3. Reducing E_R by learning global distance metric

In [17] [9], only $E_D(\omega)$ term is considered, and it is modeled as $E_D(\omega) = \| \mathbf{a} - \mathbf{A}\omega \|_2$, the Euclidean distance between reconstructed appearance and query appearance. However, as analyzed in Sec. 3.2, the space structure of the two spaces, if simply defined by Euclidean distance, will not be the same, resulting in different reconstruction weights in the two spaces.

In order to reduce the error $E_R(\omega)$ caused by this discrepancy, instead of pursuing different weights ω for the two spaces, we propose to find an optimal distance metric $\mathcal{D}^*$ for the appearance space α, such that the structure of α approximates the Euclidean structure of β, *i.e.*, $\mathcal{D}^*(\mathbf{a}_i, \mathbf{a}_j) \sim \| \mathbf{y}_i - \mathbf{y}_j \|_2^2$. We call it the global distance metric, since it is learned from the space structures defined by all training appearance-gaze pairs, and applies to all appearance.

Our method seeks the optimal distance metric $\mathcal{D}^*$:

$$\mathcal{D}^* = \arg\min_{\mathcal{D}} \epsilon(\alpha, \beta : \mathcal{D}), \tag{5}$$

where $\epsilon(\alpha, \beta : \mathcal{D})$ refers to the discrepancy between the structure of space α subject to the metric $\mathcal{D}$, and the structure of space β subject to the Euclidean distance metric. α and β are the appearance space and the gaze space, respectively. With this learned metric, our $E_D(\omega)$ term is

$$E_D(\omega) = \mathcal{D}^*(\mathbf{a}, \mathbf{A}\omega). \tag{6}$$

The $E_R(\omega)$ term has been reduced through the process of learning distance metric $\mathcal{D}^*$. The weight to reconstruct gaze is found through:

$$\begin{aligned} \omega^* &= \arg\min_{\omega} \mathcal{D}^*(\mathbf{a}, \mathbf{A}\omega) \\ s.t\ &\mathbf{1}^T\omega = 1 \end{aligned} \tag{7}$$

where $\mathbf{A} = \{\mathbf{a}_1, \dots, \mathbf{a}_k\}$ is the k-NN of $\mathbf{a}$ with respect to the distance metric $\mathcal{D}^*$. This is our first proposed approach, and the detail of the learning of this metric is described in Sec. 3.3.1 and the gaze estimation is given in Sec. 3.3.2.

3.3.1 Learning a global distance metric

To obtain the distance metric, we propose to learn a linear projection $\mathbf{C}$ that transforms the appearance space, such that the transformed appearance space shares the same Euclidean structure of the gaze space. The distance metric of the transformed appearance space is

$$\begin{aligned} \mathcal{D}(\mathbf{a}_i, \mathbf{a}_j) &= (\mathbf{C}\mathbf{a}_i - \mathbf{C}\mathbf{a}_j)^T(\mathbf{C}\mathbf{a}_i - \mathbf{C}\mathbf{a}_j) \\ &= (\mathbf{a}_i - \mathbf{a}_j)^T\mathbf{S}(\mathbf{a}_i - \mathbf{a}_j) \end{aligned} \tag{8}$$

where $\mathbf{S} = \mathbf{C}^T\mathbf{C}$ is the transformation kernel. $\mathbf{S}$ is a positive semidefinite (PSD) matrix. This is equivalent to learning a Mahalanobis distance metric with kernel $\mathbf{S}$ for the original eye appearance space.

We model the structure of the appearance space α, the gaze space β, and the transformed appearance space γ. Then we need to learn a transform kernel $\mathbf{S} : \alpha \mapsto \gamma$, such that the structure of space γ approximates the structure of space β. This structure representation can have various choices, as long as it conveys the distance information between different data pairs. Without losing generality, in this paper we represent the structure of a space by its affinity matrix.

Given the training set of appearance-gaze exemplar pairs $(\mathbf{a}_i, \mathbf{y}_i),\ i = 1 \dots N$, the affinity matrix of the target space β is $\mathbf{U} \triangleq [u_{ij}]$

$$u_{ij} = \exp(-\frac{(\mathbf{y}_i - \mathbf{y}_j)^T(\mathbf{y}_i - \mathbf{y}_j)}{2\sigma_1}). \tag{9}$$

Since the scale of the distance among data points does not change the estimation, normalization is performed on u_{ij} and we get $\mathbf{P} \triangleq [p_{ij}]$

$$p_{ij} = \frac{u_{ij}}{\sum_{k \neq i} u_{ik}},\ \ p_{ii} = 0. \tag{10}$$

The affinity matrix of transformed appearance space γ is modeled in the same way. The affinity matrix $\mathbf{V}^{\mathbf{S}} \triangleq [v_{ij}^{\mathbf{S}}]$

$$v_{ij}^{\mathbf{S}} = \exp(-\frac{(\mathbf{a}_i - \mathbf{a}_j)^T\mathbf{S}(\mathbf{a}_i - \mathbf{a}_j)}{2\sigma_2}). \tag{11}$$

By normalizing matrix $\mathbf{V}^{\mathbf{S}}$, we get matrix $\mathbf{Q}^{\mathbf{S}} \triangleq [q_{ij}^{\mathbf{S}}]$

$$q_{ij}^{\mathbf{S}} = \frac{v_{ij}^{\mathbf{S}}}{\sum_{k \neq i} v_{ik}^{\mathbf{S}}},\ \ q_{ii}^{\mathbf{S}} = 0. \tag{12}$$

We use KL divergence to model this difference between the matrices $\mathbf{P}$ and $\mathbf{Q}^{\mathbf{S}}$.

$$KL[\mathbf{P}|\mathbf{Q}^{\mathbf{S}}] = \sum_{ij} KL[p_{ij}|q_{ij}^{\mathbf{S}}]. \tag{13}$$

Hence, the objective function of learning the metric is:

$$\arg\min_{\mathbf{S}} f(\mathbf{S}) = \sum_{ij} KL[p_{ij}|q_{ij}^{\mathbf{S}}],\ \ \mathbf{S} \in PSD \tag{14}$$

Equ. 14 can be solved by alternating between gradient descent step and projection to PSD cone. Noticing that:

$$f(\mathbf{S}) = \sum_{ij} p_{ij} \log p_{ij} - \sum_{ij} p_{ij} \log q_{ij}^{\mathbf{S}} \tag{15}$$

The gradient of $f(\mathbf{S})$ w.r.t. $\mathbf{S}$ is

$$\nabla f(\mathbf{S}) = \frac{1}{2\sigma_2} \sum_{ij} (p_{ij} - q_{ij}^{\mathbf{S}})(\mathbf{a}_i - \mathbf{a}_j)(\mathbf{a}_i - \mathbf{a}_j)^T \tag{16}$$

For t-th iteration, gradient descent is performed with step length η to get $\mathbf{S}$ of next iteration.

$$\mathbf{S}^{t+1} = \mathbf{S}^t - \eta\nabla f(\mathbf{S}) \tag{17}$$

To ensure that $\mathbf{S}$ is PSD, we project the matrix $\mathbf{S}^{t+1}$ to PSD cone. We first perform EVD on $\mathbf{S}^{t+1}$

$$\mathbf{S}^{t+1} = \sum_k \lambda_k \mathbf{u}_k \mathbf{u}_k^T \tag{18}$$

where λ_k is the eigenvalue of matrix $\mathbf{S}$, $\mathbf{u}_k$ is its corresponding eigenvector. Then we eliminate the components corresponding to negative eigenvalue.

$$\mathbf{S}^{t+1} = \sum_k \max(0, \lambda_k) \mathbf{u}_k \mathbf{u}_k^T \tag{19}$$

Alternate the process of gradient descent and projection on PSD cone until $\mathbf{S}$ converges.

Compared with similar metric learning methods in [5] [4], the proposed method is different in following aspects. Firstly, the proposed method is for regression problems, rather than classification. Secondly, our objective is to align the structure of the two spaces (*i.e.*, the visual appearance space and the gaze space), not to collapse all the feature points in the same class to a single point.

3.3.2 Gaze estimation with learned distance metric

With the transformation kernel $\mathbf{S} = \mathbf{C}^T\mathbf{C}$, the new objective of gaze reconstruction corresponding to Eq. 7 is:

$$\begin{aligned} &\arg\min_{\omega} \parallel \mathbf{a} - \mathbf{A}\omega \parallel_{\mathbf{C}}^2 \\ &s.t. \mathbf{1}^T\omega = 1, \end{aligned} \tag{20}$$

where $\mathbf{a}$ is the input query appearance feature, $\mathbf{C}$ is the transform matrix, and $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_k)$ is the collection of the k-nearest neighbors of the query appearance. Note that here the k-nearest neighbors are chosen based on the Mahalanobis distance with kernel $\mathbf{S}$.

It is easy to obtain the closed-form solution to this constrained least squares problem:

$$\omega^* = \frac{\mathbf{D}^\dagger \mathbf{1}}{\mathbf{1}^T \mathbf{D}^\dagger \mathbf{1}} \tag{21}$$

where $\mathbf{D} = (\mathbf{a}\mathbf{1}^T - \mathbf{A})^T \mathbf{S} (\mathbf{a}\mathbf{1}^T - \mathbf{A})$, $\mathbf{D}^\dagger$ is the pseudoinverse of matrix $\mathbf{D}$.

3.4. Reducing E_R by local regularization

In Sec. 3.3, the reduction of E_R term is conducted in training phase with a global scope. In this section, we consider the possibility to further decrease E_R in the estimation phase with a local scope.

In the estimation part in Sec. 3.3.2, Equ. 20 minimizes the matching error between the reconstructed appearance $\mathbf{A}\omega$ and the query appearance $\mathbf{a}$ with the learned distance metric $\mathcal{D}^*$. It only considers the space structure of the training set of appearance and gaze pairs $(\mathbf{a}_i, \mathbf{y}_i), i = 1 \dots N$. Nevertheless, when the gaze $\mathbf{y}$ corresponding to the query appearance $\mathbf{a}$ is estimated (*i.e.*, not included in the training set), we actually obtain another pair of appearance and gaze, $(\mathbf{a}, \mathbf{B}\omega)$, where $\mathbf{a}$ is the query appearance, $\mathbf{B}\omega$ is the reconstructed or estimated gaze position. Therefore, the optimal weight ω^* has impact on the mutual information between the true gaze and the estimated gaze $\mathbf{B}\omega$. However, this is missing in the estimation in Equ. 20, which will introduce additional error part of $E_R(\omega)$ to the final estimation.

This part of $E_R(\omega)$ can be further reduced in the estimation phase. We denote by β' the reconstructed gaze space (*i.e.*, via $\mathbf{B}\omega$). Note that when $\epsilon(\alpha, \beta : \mathcal{D}^*)$ is small, $\epsilon(\alpha, \beta' : \mathcal{D}^*)$, the discrepancy of the structure of the appearance space α using distance metric $\mathcal{D}^*$ and space β' with Euclidean distance, should also be small. We use this constraint to further regularize this space structure discrepancy.

This $E_R(\omega) = \epsilon(\alpha, \beta' : \mathcal{D}^*)$ term is local because it only considers the local space structure, *i.e.* only the local data related to the query appearance $\mathbf{a}$ and the estimated gaze $\mathbf{B}\omega$ is influenced by the estimated ω. Only this local space structure discrepancy will contribute to $E_R(\omega)$. By penalizing this local $E_R(\omega)$ term, ω is not only trying to find the best matched reconstructed appearance $\mathbf{A}\omega$, but also trying to minimize the error caused in reconstructing the estimated gaze $\mathbf{B}\omega$ by using this ω. $\epsilon(\alpha, \beta' : \mathcal{D}^*)$ is analogous to the energy function in energy-based classification in [19], but with different objective function. $\epsilon(\alpha, \beta' : \mathcal{D}^*)$ denotes the discrepancy between the structure of two spaces, while the energy term in [19] is hinge loss.

3.4.1 Modeling the local space structure

To describe the structure of extended gaze space, we use

$$u_i = \exp(-\frac{(\mathbf{B}\omega - \mathbf{y}_i)^T(\mathbf{B}\omega - \mathbf{y}_i)}{2\sigma_1}), \tag{22}$$

where $\mathbf{B}\omega$ is the estimated gaze, $\mathbf{B} = \{\mathbf{y}_1, \dots, \mathbf{y}_k\}$ that is the collection of the corresponding gazes of the k-nearest neighbor of the query appearance. We normalize it:

$$p_i = \frac{u_i}{\sum_k u_k} \tag{23}$$

The extended appearance space is subject to the Mahalanobis kernel $\mathbf{S}$, and its structure can be modeled as:

$$v_i^{\mathbf{S}} = \exp(-\frac{(\mathbf{a} - \mathbf{a}_i)^T \mathbf{S} (\mathbf{a} - \mathbf{a}_i)}{2\sigma_2}). \tag{24}$$

After normalization, we have:

$$q_i^{\mathbf{S}} = \frac{v_i^{\mathbf{S}}}{\sum_k v_k^{\mathbf{S}}}. \tag{25}$$

The discrepancy of the local structure of spaces α and β' can still be modeled by the KL divergence $KL[\mathbf{p}|\mathbf{q}^{\mathbf{S}}]$, where vector $\mathbf{q}^{\mathbf{S}} \triangleq [q_i^{\mathbf{S}}]$ models the structure of appearance space α related to the query appearance $\mathbf{a}$, vector $\mathbf{p} \triangleq [p_i]$ models the structure of the gaze position space β' related to the estimated gaze position $\mathbf{B}\omega$. Similar to Sec. 3.3.1,

$$KL[\mathbf{p}|\mathbf{q}^{\mathbf{S}}] = \sum_i p_i \log p_i - \sum_i p_i \log q_i^{\mathbf{S}}. \tag{26}$$

3.4.2 Estimation with local regularization

The final objective function of our proposed method is:

$$\begin{aligned} \arg\min_{\omega} f(\omega) &= \| \mathbf{a} - \mathbf{A}\omega \|_{\mathbf{C}}^2 + \lambda\, KL[\mathbf{p}|\mathbf{q}^{\mathbf{S}}] \\ s.t. \mathbf{1}^T \omega &= 1 \end{aligned} \tag{27}$$

where λ is a balancing factor.

$$f(\omega) = \| \mathbf{a} - \mathbf{A}\omega \|_{\mathbf{C}}^2 + \lambda\Big(\sum_i p_i \log p_i - \sum_i p_i \log q_i^{\mathbf{S}}\Big), \tag{28}$$

whose gradient w.r.t. ω is:

$$-2\mathbf{A}^T\mathbf{S}(\mathbf{a}-\mathbf{A}\omega)+\frac{\lambda}{\sigma_1}\Big(\sum_i p_i \log \frac{p_i}{q_i^{\mathbf{S}}}(\mathbf{B}^T\mathbf{y}_i-\sum_j p_j \mathbf{B}^T\mathbf{y}_j)\Big). \tag{29}$$

Gradient-based method can be easily applied.

4. Experiments

This section evaluates the performance of the proposed method. Experiments have been performed on the benchmark dataset provided by Sugano *et al.* in [15]. Same as in [9], the estimation error is measured in degree, and calculated as

$$Error = \arctan\Big(\frac{\| \hat{\mathbf{y}} - \mathbf{y} \|_2}{d}\Big), \tag{30}$$

where $\hat{\mathbf{y}}$ is the estimated gaze position, $\mathbf{y}$ is the ground truth gaze position, and d is the distance between experiment subject and the screen on which the gaze position resides.

Baseline comparisons include k-NN, SVR, Local Linear Reconstruction (LLR), the triangle region based Local Linear Reconstruction (LLR-TRI) [17] and Adaptive Linear Regression (ALR) [9]. The LLR method is a simplified version of [17], where the k-NN is chosen only based on the Euclidean distance without considering the topological information as in LLR-TRI. To investigate the influence of the regularization term in Sec. 3.4, the proposed method has been evaluated under two conditions: without the regularization term, denoted as *Ours* (Sec. 3.3.2), and with regularization term, denoted as *Ours-R* (Sec. 3.4.2).

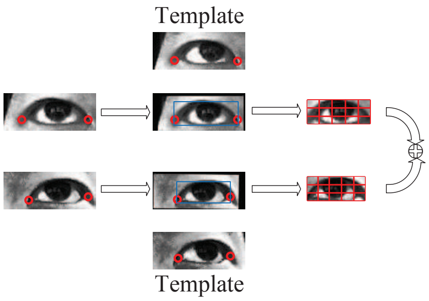


Figure 2. Illustration of feature extraction process

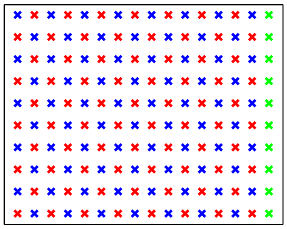


Figure 3. Gaze position patterns. Each blue cross denotes one gaze position in the training set. Each red cross is one gaze position in the testing set. Each green cross denotes one excluded gaze position in UT multi-view gaze dataset.

The proposed methods and baseline methods were evaluated in three scenarios, fixed-subject and fixed-pose setting, cross-pose setting, and cross-subject setting. Influence of the balancing parameter λ, and the dimension of appearance will be discussed in following sections.

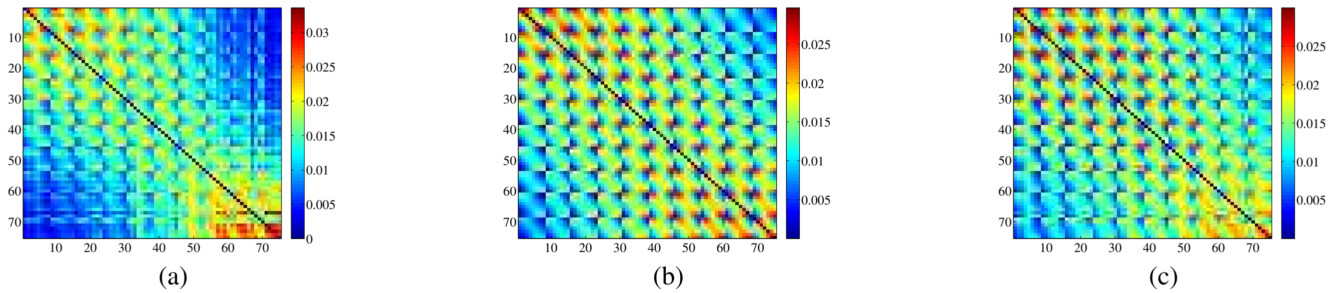
4.1. Appearance feature

We extract the eye appearance feature in following steps:

1. Histogram equalization is performed on eye images.
2. Follow method in [9] to detect the right and left eye corners, align the eye images with respect to the two eye corners to the template eye images.
3. The image region bounded by two eye corners is cropped using a fixed aspect ratio.
4. Eye images are downsampled to a lower resolution, *e.g.* 3×5, after antialiasing filtering. Normalize and concatenate appearance feature $\mathbf{a}_l$ and $\mathbf{a}_r$ of the left eye and the right eye to one appearance feature, $\mathbf{a} = (\frac{\mathbf{a}_l^T}{\|\mathbf{a}_l\|_2}, \frac{\mathbf{a}_r^T}{\|\mathbf{a}_r\|_2})^T$.

4.2. Result on UT multi-view gaze dataset

Experiments are performed on a recently published dataset, UT Multi-view Gaze Dataset [15], which contains 50 subjects, 160 gaze directions, 8 cameras and 144 synthetic head poses for each subject. The proposed methods and baseline methods were evaluated under three settings,



(a) (b) (c)

Figure 4. Result of metric preservation. Figure (a), (b), (c) shows the affinity matrix of appearance feature space, gaze position space and the transformed appearance feature space after metric learning.

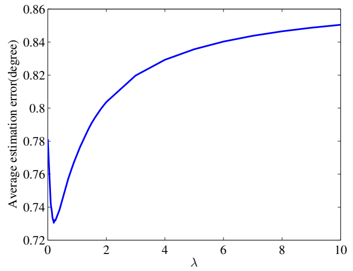


Figure 5. Regularized estimation errors with different λ

including *fixed-pose & fixed-subject, cross-pose, cross-subject*.

In fixed-pose and fixed-subject setting, for each subject, training and testing images are from the same camera. The training and testing gaze split pattern follows [9], as shown in Fig. 3. One camera is sufficient for fixed pose and subject setting. We randomly choose camera No.5. In total, there are 3750 training samples and 3750 testing samples. For each subject, there are 75 training samples and 75 testing samples. Feature extraction follows Sec. 4.1. Note that ALR method requires the feature dimension to be lower than the number of training samples, the dimension of appearance feature of each eye must be lower than 37. On the contrary, our method does not have such constraint. For a fair comparison, 30-dimensional appearance features are constructed, denoted by 30-D, with 3×5 dimension for each eye image, the same as [9]. k for k-nearest neighbor is set to 8. The dimension of the appearance feature is later increased to 64, denoted by 64-D, 4×8 for each eye image, to investigate its influence on estimation accuracy.

In cross-pose setting, we use synthesized images of each subject for training, and actual recorded images of the same subject for testing. The testing data has 8 different poses (cameras). All of 160 gaze positions of each pose are included in corresponding training or testing set, with no split. k for k-NN is increased to 10 as more training data are included. Note that the synthesized poses may also include the actual poses, we exclude from training set the synthetic poses whose distance to the testing pose is less than 5 degrees. This setting is not investigated in [15] as they use all synthetic poses as training. Note that because of normalization, the poses of left eye and right eye are different for the same actual camera. Therefore, we can not use appearance which combines two eye images as in Sec. 4.1 since training poses are chosen based on testing poses. Left eye and right eye were evaluated separately. Appearance feature is extracted from each eye and with 9×15 dimension as in [15].

Cross-subject setting follows the design in [15]. Training images include all gaze positions of 144 synthesized poses of 33 different subjects. Testing images include all 160 gaze positions of 8 actual cameras. k of k-NN is set to 10. Appearance feature is single eye image of 9×15 dimension.

4.2.1 Metric learning result

Fig. 4 shows the affinity structure of the original appearance space, gaze space and the transformed appearance space of subject 00 with fixed-pose and fixed-subject setting. $\mathbf{S}$ is initialized as identity matrix. It is clear that the structure of the original appearance space in Fig. 4 (a) is deviated from the gaze space in Fig. 4 (b). While the gaze space exhibits a strong periodic structure, the original appearance feature space structure is cluttered and irregular. As shown in Fig. 4 (c), it is clear that the affinity structure of the transformed appearance feature space better approximates the structure of the gaze space.

4.2.2 Influence of λ

Fig. 5 shows how the average estimation error changes as the balancing factor λ varies, under fixed-pose and fixed-subject setting. When λ is 0, the estimation error is large since the objective function degenerates to Equ. 20, which only considers the reconstruction error with learned distance metric but overlooking $E_R(\omega)$. When λ is too large, the estimation is also inferior since the $E_D(\omega)$ may be large, *i.e.* the reconstructed appearance is deviated from the query

	Error 30-D	Error 64-D
k-NN	2.49±1.84	2.47±1.80
SVR	1.32±1.32	1.05±1.26
LLR	1.23±1.34	1.11±1.24
LLR-TRI	1.27±1.60	1.14±1.53
ALR	1.23±1.35	1.06±1.75
Ours	1.14±1.20	0.95±1.10
Ours-R	**1.08±1.15**	**0.92±1.08**

Table 1. Estimation error without pose and subject variation.

	Cross-pose	Cross-Subject
k-NN	4.27±2.95	7.51±3.99
SVR	6.46±3.76	7.57±4.70
LLR	4.31±2.96	7.93±4.33
LLR-TRI	8.37±5.10	8.86±5.12
ALR	7.28±4.42	7.75±4.17
Ours	3.63±2.56	6.34±3.72
Ours-R	**3.39±2.47**	**6.14±3.69**

Table 2. Estimation error with pose and subject variation.

one, so is the estimated gaze. As shown in Fig. 5, the curve has only one extreme point, which makes it easy to find the optimal λ via binary search and cross validation on the same training set. In fixed-pose and fixed-subject setting, λ is fixed for different gaze directions. In cross-pose setting and cross-subject setting, it is estimated from k-NN training features of the query feature, which will be introduced in Sec. 4.2.4.

4.2.3 Result without pose and subject variation

This section investigates the fixed-pose and fixed-subject setting. Average estimation error in degree and standard deviation of error have been calculated for evaluation. To assess statistical significance of our proposed method, the Wilcoxon signed rank test has been conducted between each baseline method and *Ours-R*. A *p* value less than 0.05 indicates that the difference between the estimation error by *Ours-R* and the estimation error by one baseline method is statistically significant.

The average estimation error across all subjects is summarized in table 1 in the form of $mean \pm std$. The best estimation error is marked by bold face, while the second best one is in blue color. With 30-D appearance feature, in average of all subjects, *Ours-R* outperforms baseline methods relatively by 11.7% to 56.6%, *Ours* by 5.2%. With 64-D appearance feature, *Ours-R* outperforms baseline methods relatively by 11.8% to 62.7%, *Ours* by 3.2%. *Ours-R* also achieves the lowest standard deviation. In average of all subjects, *p* value of each baseline is less than 0.05

This experimental result shows that as the feature dimension increases, the average estimation error for all methods decreases. With 64-D appearance feature, our proposed methods still outperforms baseline methods by a significant margin similar to the 30-D feature. Moreover, since ALR method requires appearances with dimension lower than the number of training exemplars, it cannot use the higher dimensional feature, while our proposed methods can.

4.2.4 Result with pose and subject variation

In this section, we show that the proposed method is actually robust to pose and subject variations. With pose and subject variations, the visual appearance space structure is more complex than that under fixed-pose and fixed-subject setting. A simple linear distance metric may not be able to align the entire appearance space to the gaze space. Our method indeed performs both the global alignment (*i.e.*, learning a global metric) and local alignments conditioned on the query (*i.e.*, a local metric). At the offline training, a rough global metric is learned on the entire appearance space. Then for a given query, we further learn a fine metric using the k-NN of this query, where k-NN exemplars are selected based on the global metric, as described in Sec. 3.3.1. λ is estimated using these k-NN exemplars.

Experimental results are summarized in table 2. *Ours-R* achieves the best estimation error with *p* values less than 0.05. Our methods produce obvious performance improvement over LLR. It clearly shows that the proposed method is resilient to the head pose and subject variations. In addition to average cross-pose error, we also investigate how the estimation error changes w.r.t. the distance between training and testing poses. Due to limited space, the result is provided in supplementary material.

5. Conclusion

This paper investigates a fundamental issue in reconstruction-based gaze estimation, *i.e.*, under what condition the reconstruction in the appearance space can be transferred to the gaze space for gaze estimation. This is the first of its kind to study this important issue. Without a proper metric adjustment in either space, the direct transfer of the reconstruction from one space to the other is questionable. This paper proposes an effective metric learning method to identify a new metric for the appearance space. In addition, this paper also presents a novel method to solve for the reconstruction by local regularization based on the affinity structure of the appearance space and the reconstructed gaze space. These two methods have shown consistently and significantly superior performances on public benchmark dataset across different settings.

6. Acknowledgement

This work was supported in part by National Science Foundation grant IIS-1217302 and ARO W911NF-15-1-0472.

References

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

[2] J. Chen and Q. Ji. 3d gaze estimation with a single camera without ir illumination. In *ICPR*, pages 1–4, 2008. 2

[3] J. Chen and Q. Ji. Probabilistic gaze estimation without active personal calibration. In *CVPR*, pages 609–616, 2011. 2

[4] A. Globerson and S. T. Roweis. Metric learning by collapsing classes. In *NIPS*, pages 451–458, 2005. 5

[5] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*, pages 513–520, 2004. 5

[6] E. D. Guestrin and E. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6):1124–1133, 2006. 2

[7] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *PAMI*, 32(3):478–500, 2010. 2

[8] F. Lu, T. Okabe, Y. Sugano, and Y. Sato. A head pose-free approach for appearance-based gaze estimation. In *BMVC*, pages 1–11, 2011. 1, 2

[9] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. In *ICCV*, pages 153–160, 2011. 1, 2, 3, 6, 7

[10] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Head pose-free appearance-based gaze sensing via eye image synthesis. In *ICPR*, pages 1008–1011, 2012. 1, 2

[11] K. A. F. Mora and J.-M. Odobez. Geometric generative gaze estimation (g3e) for remote rgb-d cameras. In *CVPR*, pages 1773–1780, 2014. 2

[12] C. H. Morimoto and M. R. Mimica. Eye gaze tracking techniques for interactive applications. *CVIU*, 98(1):4–24, 2005. 1, 2

[13] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 2

[14] Y. Sugano, Y. Matsushita, and Y. Sato. Appearance-based gaze estimation using visual saliency. *PAMI*, 35(2):329–341, 2013. 2

[15] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *CVPR*, 2014. 2, 6, 7

[16] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike. An incremental learning method for unconstrained gaze estimation. In *ECCV*, pages 656–667, 2008. 1, 2

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

[18] J. Wang, E. Sung, and R. Venkateswarlu. Eye gaze estimation from a single image of one eye. In *ICCV*, pages 136–143, 2003. 2

[19] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2005. 5

[20] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the s^3gp. In *CVPR*, volume 1, pages 230–237, 2006. 2

[21] L.-Q. Xu, D. Machin, and P. Sheppard. A novel approach to real-time non-intrusive gaze finding. In *BMVC*, pages 1–10, 1998. 2

[22] D. H. Yoo and M. J. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *CVIU*, 98(1):25–51, 2005. 2

[23] D. H. Yoo, J. H. Kim, B. R. Lee, and M. J. Chung. Non-contact eye gaze tracking system by mapping of corneal reflections. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 94–99, 2002. 2

[24] Z. Zhu and Q. Ji. Eye gaze tracking under natural head movements. In *CVPR*, volume 1, pages 918–923, 2005. 2

[25] Z. Zhu, Q. Ji, and K. P. Bennett. Nonlinear eye gaze mapping function estimation via support vector regression. In *ICPR*, volume 1, pages 1132–1135, 2006. 2