Neurocomputing 182 (2016) 10–17

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Person-independent eye gaze prediction from eye images using patch-based features

Feng Lu [a,b], Xiaowu Chen [c,*]

[a] School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[b] International Research Institute for Multidisciplinary Science, Beihang University, Beijing 100191, China
[c] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form
1 June 2015
Accepted 18 July 2015
Communicated by Ke Lu
Available online 18 December 2015

Keywords:
Gaze estimation
Eye image
Sparse auto-encoder
Gaze direction classification

ABSTRACT

This paper delivers a preliminary attempt towards person-independent appearance-based gaze estimation. Conventional methods need to assume training and test data collected from the same person, otherwise eye shape difference due to individuality will affect the estimation severely. To solve this problem, the key idea in this paper is to extract from eye images more advanced eye features, which helps learn a person-independent relationship between eye gaze change and eye appearance variation. To this end, we propose employing the advantages of recent sparse auto-encoding techniques. We partition any eye image into small patches which can overlap with each other. With patches from many images, we learn a codebook comprising a set of bases, which can reconstruct any eye image patch with sparse coefficients. By examining these coefficients, we can analyze the eye shape more effectively. Finally, we produce the eye features by pooling the coefficients at different scales, and then combine these subfeatures from different codebooks. Experimental results show that the proposed method achieves good accuracy on a public dataset and it also outperforms conventional methods by a large margin.

1. Introduction

Human gaze tracking has been long considered as the next generation solution for human computer interaction (HCI) [1]. It is supposed to be much more powerful than any other conventional interfaces. The reason is that more than half of human sensory information is received by the eyes and processed by the human visual systems in the brain. Therefore, eye gaze movements can be used as an essential cue to directly reflect human attention, feeling, intention and other internal status [2,3], which are important for the design of an effective interaction interface for applications including virtual reality (VR), HCI and many other promising systems [4–6].

However, in the field of computer vision, existing gaze estimation techniques still face various difficulties and limitations. The so-called model-based methods, which occupy the majority of existing solutions, require to extract small eye features in the eye image. To this end, they usually need special devices such as infrared lights and multiple cameras. On the other hand, another type of methods, namely appearance-based methods, can use the entire eye image as input to learn a simple regression. They can work with low resolution images without active illuminations. However, training is necessary every time before use. In fact, the appearance-based methods only accept training data and test data from the same person. Otherwise the eye shape change due to individuality will greatly affect the accuracy.

In this paper, we propose a method that belongs to the category of appearance-based methods. Our focus is to develop a technique that can use training data from other persons when doing gaze estimation for the current person. In this sense, training can be done in advance without the participation for the current user. In other words, a user can use a pre-trained system directly without performing troublesome training every time. This makes a huge difference compared to conventional appearance-based methods.

To solve this problem, our idea is to propose a better eye feature that is less person-dependent during training and test. In particular, inspired by the recent developments on deep network and various autoencoders for image feature learning [7,8], we partition eye images into patches and use a sparse auto-encoder to learn a set of bases from randomly collected patches. When we use the learnt bases to reconstruct any original image patch, the resulting coefficients only have sparsely distributed non-zero values. In other words, every non-zero coefficient plays an important role in carrying essential structural information of the

* Corresponding author.
E-mail addresses: lufeng@ut-vision.org (F. Lu), chen@buaa.edu.cn (X. Chen).

http://dx.doi.org/10.1016/j.neucom.2015.07.125
0925-2312/

original image patch. Therefore, the obtained coefficients have a good potential to be used as eye features in our problem. Furthermore, we propose to obtain coefficients by using multiple codebooks, and use spatial pyramid pooling to extract final features from the coefficients at different spatial layers. The resulting feature vector is able to handle factors such as scaling, translation and even shape deformation well, and thus it greatly improves the gaze estimation accuracy without training by the same person.

Overall, this paper makes the following contributions: (1) computation of patch-based sparse representation of the eye appearance (Section 3); (2) final eye feature by using multiple codebooks and spatial pyramid pooling (Section 4.1); (3) modeling and solving the problem by using the proposed eye feature (Section 4.2). Experimental evaluations in Section 5 demonstrate the advantage of the proposed method.

2. Related works

There have been many computer vision-based gaze estimation techniques proposed recently. According to recent surveys [9,10], most existing methods belong to either of the two major categories, namely the model-based methods and appearance-based methods.

The model-based methods assume some kinds of eyeball models, based on which eye gaze directions can be computed geometrically. Such models can be either 2D or 3D [11,12], or can be other models that involve environment configurations, e.g., cross ratio models [13,14]. In order to obtain the key parameters of the model, small features on the eyeball surface are extracted from eye images. The most commonly used features include near infrared (NIR) corneal reflections, pupil reflections [13,11,15] and iris contours [16,17]. To produce the NIR reflection points on the eyeball surface, active NIR illumination is usually assumed, together with infrared cameras with long focal lengths and multiview cameras for eyeball position tracking. Besides, positions of lights and cameras are always calibrated beforehand [18–20].

Recently, several methods have been proposed leveraging the depth information from depth sensors, e.g., Kinect [21–23]. The depth information is very useful in tracking user's head pose and locating the eye region. However, a depth sensor is also considered to be additional hardware with active illuminations and capturing systems. Similar to conventional model-based methods, they are also less practical in some common scenarios where only a mobile phone or tablet is being used.

Most appearance-based methods do not extract small eye features. Instead, they use all pixel values from an entire eye image as a high dimensional feature vector, and learn a mapping between eye features and real gaze positions through training. Such a mapping can be learnt via different techniques. For instance, early systems used neural networks to learn the mapping [24,25]. However, due to the large number of unknown weights in the network, thousands of training samples were needed to refine the neurons' connections. Later, linear regression became widely used. Tan et al. [26] proposed a simple method that interpolated the unknown gaze sample by using its nearest neighbors. This method exploits the local similarity of data in the eye appearance manifold and reduces the number of training samples to several hundreds. To further reduce the training cost, Williams et al. [27] proposed a semi-supervised method that could use both labeled and unlabeled samples to train a Gaussian Process Regressor. Lu et al. [28,29] followed the idea of linear regression and introduced an adaptive regression method to use much sparsely collected training samples. Their method can also handle problems such as alignment and eye blink in gaze estimation. Sugano et al. [30] proposed a calibration-free method by assuming that a user is watching a video. Visual saliency information from the video can then guide the gaze position prediction. However, when free head motion is considered, most existing methods need to increase their training samples' number [31–34] again.

Another common limitation shared by existing appearance-based methods is that they all assume that training data and test data are collected during the same session, i.e., from the same user. Therefore, a training stage becomes a must before a user can use the system. In contrast, if a system can be trained by someone else beforehand, while it can still work well for a new user, its applicability will be dramatically improved. Delivering techniques towards such a goal is the main purpose of this work.

3. Patch-based eye feature extraction using sparse auto-encoder

In this section, we propose to obtain patch-based features from eye images using sparse auto-encoder. The features are coefficients from sparse representations of images, and they carry important structural information of the eye shape more effectively than original pixel values.

3.1. Patch-based eye image codebook

Assume we are given a set of eye images $\boldsymbol{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_n,\ldots,\mathbf{x}_N]$, where each $\mathbf{x}_n$ is a column vector stacking all pixel values from one image. We aim at finding a codebook comprising bases $\boldsymbol{B}=[\mathbf{b}_1,\ldots,\mathbf{b}_m,\ldots,\mathbf{b}_M]$ that are able to represent each image by

$$\mathbf{x}_n=\sum_{m=1}^{M} c_{m,n}\cdot\mathbf{b}_m. \tag{1}$$

We can write this in a matrix form by

$$\boldsymbol{X}=\boldsymbol{BC}. \tag{2}$$

Given images $\boldsymbol{X}$ as observations, we can optimize the codebook $\boldsymbol{B}$ by solving

$$\boldsymbol{B}=\underset{\boldsymbol{B}}{\operatorname{argmin}}\|\boldsymbol{BC}-\boldsymbol{X}\|^2, \tag{3}$$

where $\boldsymbol{C}=[\mathbf{c}_1,\ldots,\mathbf{c}_n,\ldots,\mathbf{c}_N]$ and $\mathbf{c}_n=[\ldots,c_{m,n},\ldots]^{\mathrm{T}}$.

Because Eq. (3) applies a matrix decomposition without any regular term, the solution will not be unique. A randomly obtained codebook $\boldsymbol{B}$ cannot be optimal in terms of representing $\boldsymbol{X}$ efficiently. On the other hand, the sparse auto-encoder theory [37] claims that one can assume the sparsity in the coefficients $\boldsymbol{C}$, which leads to the following problem:

$$\begin{aligned}&\boldsymbol{B}=\arg\min_{\boldsymbol{B}}\|\boldsymbol{BC}-\boldsymbol{X}\|^2+\lambda\sum_{n=1}^{N}\|\mathbf{c}_n\|_1,\\ &\text{s.t.}\quad \|\mathbf{b}_m\|<T,\quad \forall m=1,\ldots,M.\end{aligned} \tag{4}$$

Note that in Eq. (4), we use an L_1 penalty in the second term. This is because for reconstructing each image, we expect as few as possible bases are activated, i.e., with non-zero coefficients. In this sense, each of the learnt basis has a maximum ability to represent an input image. Also note that we assume any basis in $\boldsymbol{B}$ has a norm that is no larger than a threshold T. This is because if the elements in $\boldsymbol{B}$ can grow freely, $\boldsymbol{C}$ will in return become very small while Eq. (4) can still hold true. This will make the L_1 penalty less influential and thus the sparsity can no longer be ensured.

The above method can learn a codebook for eye images. However, in practice, it is more effective to learn the image patches rather than the entire image. Therefore, we randomly crop image patches with a certain size from many eye images. These patches constitute the matrix $\boldsymbol{X}$ and from which we can learn a set