Engineering 7 (2021) 777–786

Contents lists available at ScienceDirect

Engineering

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

Research
Intelligent Manufacturing—Article

# Gaze Estimation via a Differential Eyes' Appearances Network with a Reference Grid

Song Gu [a], Lihui Wang [b,*], Long He [a], Xianding He [a], Jian Wang [a]

[a] Chengdu Aeronautic Polytechnic, Chengdu 610100, China
[b] Department of Production Engineering, KTH Royal Institute of Technology, Stockholm 10044, Sweden

ARTICLE INFO

*Article history:*
Received 8 November 2019
Revised 11 June 2020
Accepted 6 August 2020
Available online 30 April 2021

*Keywords:*
Gaze estimation
Differential gaze
Siamese neural network
Cross-person evaluations
Human–robot collaboration

ABSTRACT

A person's eye gaze can effectively express that person's intentions. Thus, gaze estimation is an important approach in intelligent manufacturing to analyze a person's intentions. Many gaze estimation methods regress the direction of the gaze by analyzing images of the eyes, also known as eye patches. However, it is very difficult to construct a person-independent model that can estimate an accurate gaze direction for every person due to individual differences. In this paper, we hypothesize that the difference in the appearance of each of a person's eyes is related to the difference in the corresponding gaze directions. Based on this hypothesis, a differential eyes' appearances network (DEANet) is trained on public datasets to predict the gaze differences of pairwise eye patches belonging to the same individual. Our proposed DEANet is based on a Siamese neural network (SNNet) framework which has two identical branches. A multi-stream architecture is fed into each branch of the SNNet. Both branches of the DEANet that share the same weights extract the features of the patches; then the features are concatenated to obtain the difference of the gaze directions. Once the differential gaze model is trained, a new person's gaze direction can be estimated when a few calibrated eye patches for that person are provided. Because person-specific calibrated eye patches are involved in the testing stage, the estimation accuracy is improved. Furthermore, the problem of requiring a large amount of data when training a person-specific model is effectively avoided. A reference grid strategy is also proposed in order to select a few references as some of the DEANet's inputs directly based on the estimation values, further thereby improving the estimation accuracy. Experiments on public datasets show that our proposed approach outperforms the state-of-the-art methods.



## 1. Introduction

The eye gaze is informative in human communication. When working in a noisy shared space, people prefer to express their intentions through non-verbal behaviors such as eye gaze and gesture. The eye gaze carries a considerable amount of information that allows for task completion. A person's intention can be effectively perceived by estimating her or his gaze direction. Many researchers have investigated the "intention reading" ability based on gaze cues [1,2]. For example, in Ref. [1], a robot held a block in each of its hands, while a human controlled the robot successfully to make it give the human one of the blocks when the human gazed at the robot's hand. This experiment demonstrates that the rich information carried by eye gaze has a significant impact on collaboration. Gaze estimation has been applied in many domains, such as human–robot collaboration (HRC) [1,2], virtual reality [3], and mobile-device controllers [4]. In HRC in particular, gaze estimation systems will be adopted as an additional modality to control robots through multimodal fusion in addition to gestures, speech command, and body motion [5,6]. The addition of eye gaze will extend the scale of application in HRC and help improve the reliability of multimodal robot control.

In intelligent manufacturing, humans are part of the process loop in intelligent and flexible automation [7,8] and play important roles in collaboration with robots. The range of tasks that robots can deal with is increasing [9], and humans generally prefer to communicate with robots through natural methods. For example, it would be preferable to give orders to robots through a gesture or gaze rather than by using a remote controller. Furthermore,

* Corresponding author.
*E-mail address:* lihuiw@kth.se (L. Wang).

https://doi.org/10.1016/j.eng.2020.08.027

people are often unwilling to use invasive solutions, such as wearing special glasses [10] that can estimate their gaze direction. Instead, a camera can be installed in a nearby location to observe the operator, and the operator's gaze direction can be estimated by analyzing the digital image captured by the camera. This is a common noninvasive solution based on computer vision technology. The operator does not perceive the existence of the system when his or her gaze direction is estimated.

Noninvasive vision-based solutions can generally be divided into two types: model-based methods and appearance-based methods [11]. In model-based methods, geometric models of parts of the eye, such as the radius and the center of the pupil, are evaluated by analyzing the image, and the gaze direction is estimated based on the geometric models [12,13]. In appearance-based methods, the gaze direction is directly regressed by analyzing images of the eyes, known as eye patches. On the one hand, compared with appearance-based methods, the accuracy of the estimated direction in model-based methods depends on the quality of the captured image, such as the image resolution and illumination, because certain edges or feature points must be extracted accurately. In contrast, appearance-based methods do not require feature points. Ref. [14] evaluates popular gaze estimation methods to demonstrate that appearance-based methods achieve better performance than model-based ones. On the other hand, it is a challenging task with model-based methods to obtain a good model based on prior knowledge in order to estimate the gaze direction accurately [15]. However, deep neural networks can effectively identify the intrinsic features of the data. The successful application of deep neural networks in appearance-based methods increases the estimation accuracy dramatically. Thus, appearance-based methods have attracted a great deal of attention in recent years [16–18]. Refs. [19,20] propose video-based gaze estimation systems, which are model-based methods. It is possible to enhance the performance of the system by means of a deep neural network, such as a recurrent neural network or a long short-term memory network. However, such usage is beyond the scope of this paper.

With appearance-based methods, the key step is to determine the relationship between the input images and the gaze directions. Many researchers have constructed varying models to fit the relationship. These models are trained and tested on data from different persons, in what are referred to as cross-person evaluations. The corresponding model is denoted as a person-independent model. Because a person-independent model does not contain information about the tested person, individual differences in appearance will affect the estimation accuracy. If certain conditions from the testing process, such as the tested person's appearance, the level of illumination that will be present at the testing site, and so on, are involved when models are constructed in the training process, the system's performance will be improved. A common method is to collect labeled data belonging to the tested person for model training. This is referred to as a person-specific model. However, learning a person-specific model requires a large amount of labeled data. Collecting person-specific training data is a time-consuming task, which limits the applicability of such methods. Although some technologies, such as those discussed in Refs. [21,22], have been proposed to decrease the complexity of the collecting phase, these methods still require a great deal of training data. Inspired by Refs. [23–25], we propose that the input images and output directions be replaced with differential ones. Once the relationship between the difference of both the input images and the difference of both the gaze directions is constructed, only a few labeled images of the new person are required, and these can be treated as one of the inputs in the testing stage. Using this method, the gaze direction will also be estimated accurately.

In this paper, we propose a differential eyes' appearances network (DEANet) to estimate gaze direction based on a deep neural network learning framework. The proposed network is based on a Siamese neural network (SNNet) [26], which has two identical branches. A pair of sample-sets are fed into both of the network's branches simultaneously. Each sample-set includes both the left eye patch and the right eye patch in an image. Both patches are fed into one of the branches as a part of the multi-stream architecture [27]. The features are extracted from all the patches by each branch of the network, which contains two VGG16 networks [28] with different parameters. The outputs of both branches, in combination with the head pose information, are concatenated. The output of the network is the differential gaze of the pairwise sample-sets, followed by some full link layers. In the testing stage, a labeled sample-set belonging to the tested person, which is taken as the reference sample-set, is fed into one of the network's branches. The tested sample-set is fed into another network branch, and the output of the network is the gaze difference between the reference sample-set and the tested one. Because the gaze direction of the reference sample-set is labeled, the estimated gaze direction is equal to the network's output plus the labeled gaze direction corresponding to the reference sample-set. Moreover, a reference selection strategy can be adopted to enhance the system's performance if a few reference sample-sets are labeled. Our proposed approach assumes that the difference in the appearance of each of a person's eyes is related to the difference in the corresponding gaze directions. Because the information of the tested person is embedded into the trained models in the testing stage, the estimation accuracy is improved. Furthermore, only a few labeled images of the tested person are needed when estimating the gaze direction of that person. The proposed network does not need a large amount of data for training a person-specific model. Evaluations on many popular datasets show that our proposed algorithm performs favorably against other state-of-the-art methods.

Our contributions can be summarized as follows:

(1) This work provides a new formulation for differential gaze estimation that is integrated with both eye images and the normalized head pose information. A multi-stream architecture is fed into each of the branches in an SNNet. The SNNet-based framework not only incorporates information about the tested person in the testing stage, but also does not require the collection of a large amount of data for training a person-specific model.

(2) A reference selection strategy is provided. In this paper, a novel approach for a reference sample-set selection strategy is proposed to improve the estimation accuracy. A reference grid is constructed in the gaze space, and valid reference sample-sets are directly selected by the estimation values, which simplifies the computation of the system.

The rest of this paper is organized as follows. Related works are introduced in Section 2. Our proposed approach is then demonstrated in detail in Section 3. Experimental results and discussions are presented in Section 4. Finally, a conclusion and a future research plan are highlighted in Section 5.

## 2. Related work

This section provides a brief overview of recent works in appearance-based gaze estimation, person-specific estimation, and SSNets.

### 2.1. Appearance-based gaze estimation

Most appearance-based algorithms for gaze estimation are regarded as regressive solutions. The estimated gaze direction is a function of the input image. Intuitively, eye patches carry the greatest amount of information on the gaze direction (of the left and right eye) and should be sufficient to estimate the gaze

direction. Zhang et al. [29] proposed a method for in-the-wild appearance-based gaze estimation based on a multimodal convolutional neural network (CNN). Lian et al. [30] presented a shared CNN to estimate the gaze direction in multi-view eye patches captured from different cameras. Liu et al. [23,25] demonstrated the direct training of a differential CNN to predict the gaze difference between a pair of eye patches. Park et al. [31] proposed a novel pictorial representation in a fully convolutional framework to estimate the gaze direction. However, aside from eye patches, many other elements also affect the estimation accuracy, such as the head position, the scale of the eyes in the image, the head pose, and so forth. Some information should be embedded in the system. Liu et al. [32] used both the eye patches and an eye grid to construct a two-step training network to improve the estimation accuracy on mobile devices. Krafka et al. [4] took the eye patches, full-face patch, and the face grid as their system's input, and obtained a promising performance. Wong et al. [33] proposed a residual network model that incorporated the head pose and face grid features to estimate the gaze direction on mobile devices. In Ref. [34], the gaze was divided into three regions based on the localization of the pupil centers, and an Ize-Net network was constructed to estimate the gaze direction using an unsupervised learning technique. Yu et al. [17] introduced a constrained landmark-gaze model to achieve gaze estimation by integrating the eye landmark locations. Funes-Mora and Odobez [35] proposed a head pose invariance algorithm for gaze estimation based on RGB-D cameras and evaluated the performance on a low-resolution dataset [36]. Zhang et al. [16] analyzed the effects of all of the above information based on their own models. In Ref. [37], full-face images were used as the system's input, and an AlexNet [38] network with spatial weights was shown to significantly outperform many eye-images-input algorithms. These experiments suggest that the full-face appearance is more robust against head pose and illumination than eye-only methods. However, the full-face approach dramatically increases the computation complexity because the size of the input data is much larger than in the eye-only approach. Compression methods, such as those in Ref. [39], have been proposed in order to compress the image efficiently while preserving the estimation accuracy. It is still an open question whether the full-face approach or the eye-only approach will obtain a better performance.

Feeding raw images into the system without any pre-processing will increase the complexity of the regressive network. Some information can be normalized in the pre-processing stage in order to decrease the network's complexity. Sugano et al. [40] proposed a novel normalization method in the pre-processing stage to align the images before they were fed into the network. All kinds of data, including the images and gaze directions, were transformed into the normalized space as well. The object's scale did not need to be considered when learning or testing the network. In Ref. [40], a virtual camera was constructed by transforming or rotating the camera to a fixed position from the person's eye. The input images and gaze directions were derived in the virtual camera coordinates. Zhang et al. [41] analyzed the normalization method in detail, and extended the original normalization method to full-face images in Ref. [37].

### 2.2. Person-specific estimation

The goal of most gaze estimation algorithms is to train a person-independent model and to achieve a good cross-person evaluation performance. A person-independent model is constructed to describe the correlation between the input image and the gaze direction. However, according to the analysis proposed in Ref. [25], the difference between the visual axis and the optical axis varies for each person. A person-independent model cannot describe the correlation between the visual axis and the optical axis accurately, but a person-specific model can accurately estimate the gaze direction. A good performance of a person-specific model was demonstrated in Ref. [16], provided that there were sufficient training samples.

The collection of samples is a time-consuming task. Many methods for simplifying sample collection have been proposed in recent papers. Sugano et al. [42] proposed an incremental learning method to update the estimation parameters continuously. In Ref. [43], many kinds of data collected from different devices were fed into a single CNN composed of shared feature extraction layers and device-specific encoders/decoders. Huang et al. [22] built a supervised self-learning algorithm to train the gaze model incrementally. Moreover, the robust data validation mechanism could distinguish good training data from noise data. Lu et al. [21] also proposed an adaptive linear regression to adaptively select an optimal set of samples for training. The number of required training samples was significantly reduced, while a promising estimation accuracy remained. Although the above methods simplify the process of data collection, many labeled samples are still required to train a person-specific model. Yu et al. [44] designed a gaze redirection framework to generate large amounts of labeled data based on a few samples. Liu et al. [23] proposed a new idea for person-specific estimation based on only one eye patch. The difference in gaze direction was estimated by an SNNet according to the corresponding images as input. A few labeled samples were required in the testing stage after the SNNet was trained.

### 2.3. Siamese neural network

An SNNet was first introduced in Ref. [26] to verify the signatures written on a pen-input tablet. One of the characteristics of an SNNet is its two identical branches. Instead of a single input, a pair of inputs with the same type and different parameters are fed into the SNNet. Consequently, the output of the network is the difference of the corresponding inputs. This method has many applications in numerous fields. Venturelli et al. [24] proposed an SNNet framework to estimate the head pose in the training stage. A differential item was added to the loss function in order to improve the learning of the regressive network. Veges et al. [45] introduced a Siamese architecture to reduce the need for data augmentation in three-dimensional (3D) human pose estimation. The closest works to ours are Refs. [23,25]. However, the SNNet proposed in Refs. [23,25] does not consider the influences of both the eyes and the head pose. Moreover, it was demonstrated in both algorithms that the reference samples affected the estimation accuracy. However, the reference selection strategy was not discussed systematically in Refs. [23,25]. It should be noted that pairwise input will dramatically increase the number of pairwise training samples. The selection of a subset in training samples is analyzed in Refs. [46–48].

## 3. Differential eyes' appearances network

Although our proposed model is a person-independent model, person-specific information will be involved in the testing stage. The system's framework is illustrated in Fig. 1. Generally speaking, the whole framework is based on an SNNet. Instead of a single input, a Siamese pair of inputs is fed to both branches in the network, respectively. Moreover, both branches share the same weights. A tested face image and a reference face image are adopted as the system's raw inputs. Each image is normalized into a left eye patch and a right eye patch by the original head pose information, $\tilde{H}$. All normalized patches are included in the Siamese pair of inputs, which are referred to as a reference sample-set $P_f$

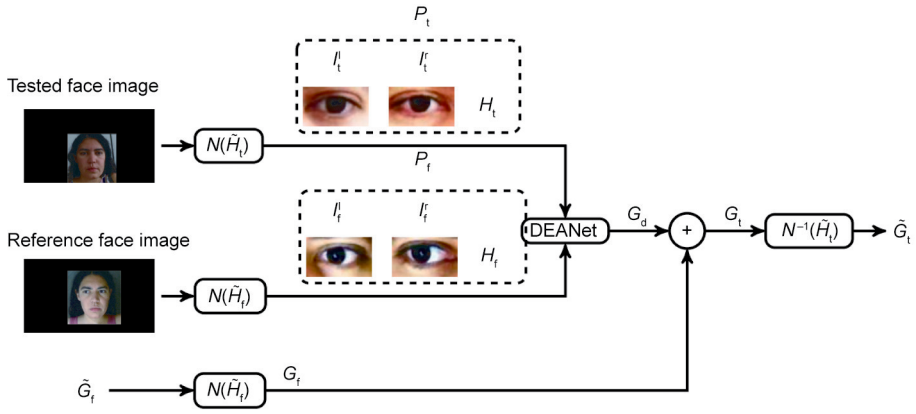


**Fig. 1.** The structure of our proposed framework. Both the tested face image and the reference face image are normalized by their original head pose information, respectively, constructing the Siamese pairs $P_t$ and $P_f$. Each Siamese pair includes a left eye patch $I^l$, a right eye patch $I^r$, and the normalized head pose information $H$, where $P_t = \{I^l_t, I^r_t, H_t\}$ and $P_f = \{I^l_f, I^r_f, H_f\}$. $G_t$ is the normalized testing gaze. The original reference gaze $\tilde{G}_f$ is labeled and then normalized, and the normalized reference gaze is denoted as $G_f$. Both Siamese pairs are fed into the DEANet to regress the differential gaze between $P_t$ and $P_f$. This is denoted as $G_d$. $N\left(\tilde{H}_t\right)$ and $N\left(\tilde{H}_f\right)$ are the same normalizing operation with different parameters. $N^{-1}\left(\tilde{H}_t\right)$, which is referred to as denormalization, is the inverse operation of the normalization with the same parameters as $N\left(\tilde{H}_t\right)$.

and a tested sample-set $P_t$, respectively. Each sample-set includes a left eye patch $I^l$, a right eye patch $I^r$, and the normalized head pose information $H$. The gaze direction corresponding to the reference data is labeled, and is referred to as reference gaze $\tilde{G}_f$. The system's output, referred to as testing gaze $\tilde{G}_t$, is the gaze direction that corresponds to the tested data. All images and $\tilde{G}_f$ are normalized by their original head pose information. Because different original head pose information will be used for the tested face image and the reference face image during normalization, it is denoted as $N\left(\tilde{H}_t\right)$ and $N\left(\tilde{H}_f\right)$, respectively, in Fig. 1. All patches that are aligned by normalization are fed into the DEANet. The normalized testing gaze is the sum of the differential gaze and the normalized reference gaze, followed by a denormalization stage, $N^{-1}\left(\tilde{H}_t\right)$, which is the inverse operation of the normalization with the same parameters as $N\left(\tilde{H}_t\right)$.

### 3.1. Definitions

The representation of the estimated direction can be categorized into two groups: two-dimensional (2D) and 3D representations. The 2D gaze position is always represented by the coordinates of the on-screen gaze location, and is used in the controller of mobile devices. The 3D gaze is a direction from the reference point to the target point in 3D space. It is composed of three angles in the camera coordinate system: the yaw, pitch, and roll. In practice, the 3D gaze is defined as a unit vector from the reference point to the target. It can then be simplified by a spherical coordinate system including $\phi$ and $\theta$; that is, $G = [\phi^g, \theta^g]'$ in this paper. Moreover, the reference point is defined as the center of the eyes. Specifically, only the 3D gaze is evaluated, and the 3D gaze direction is defined in this paper as a vector from the center of the left eye to the target. It is noted that the 2D gaze position can be derived from the 3D gaze direction when the screen plane is obtained in the 3D space. Similarly, the head pose information has the same definition as the 3D gaze direction; that is, $H = \left[\phi^h, \theta^h\right]'$ in this paper.

### 3.2. Pre-processing and normalization

As proposed in Refs. [37,40], the raw images should be normalized for gaze estimation in order to alleviate the influences caused by different cameras and the original head pose information, thereby decreasing the network's complexity. The normalization process is a series of perspective transformations so that the normalized patch is the same as the picture captured from a virtual camera looking at the same reference point. The normalization procedure and the performance have been demonstrated in detail in Refs. [40,41]. Some key steps are introduced in this section.

Initially, a single face image like the tested face image in Fig. 1 is provided. Facial landmarks, such as the corner points of the eyes and mouth, are detected by popular algorithms [49]. A left eye center point, a right eye center point, and a mouth center point, which are computed by the corner points, are used to construct a plane. The line from the right eye center to the left eye center is the $x$-axis, and the $y$-axis is perpendicular to the $x$-axis inside the plane, pointing from the eyes to the mouth. The $z$-axis is the norm of the plane conforming to the right hand rule. Integrating with the left eye center or right eye center as the original points, the three axes construct the normalized space of both eyes. According to the detected landmarks and the generic mean facial shape model [16], the normalized head pose information can then be computed by the efficient perspective-$n$-point (EP$n$P) algorithm [50]. It should be noted that both the original head pose information and the camera's intrinsic parameters are provided by popular datasets, whose performances are evaluated in Section 4. All patches that are fed into the DEANet are normalized in the normalized space. After normalization, a histogram equalization is used for all normalized patches in order to alleviate the influences caused by illumination.

The DEANet has two advantages for normalization.

(1) Normalization, as an image aligning operation, decreases the network's complexity, alleviating the influences on the eye patches caused by different camera distances, different camera intrinsic parameters, and different original head pose information. Normalized images can be simultaneously fed into the Siamese network whose branches share the same weights.

(2) Normalization simplifies the computation of the differential gaze. All parameters are in the normalized space, and the computation of gaze difference is equivalent to the operation of both gaze vectors, regardless of coordinate transformation. The proposed reference selection strategy is demonstrated for simplification in Section 3.4.

### 3.3. Training phase of the DEANet

After normalization, all patches are aligned in the normalized space regardless of the camera's intrinsic parameters and the size of the images. The normalized patches are fed into the network

to improve the system's performance because they make the network learning more efficient than un-normalized ones. Our hypothesis is that the difference in the appearance of each of a person's eyes is related to the difference in the corresponding gaze directions. Moreover, this correlation is independent of the person. To this end, a DEANet is proposed based on an SNNet for appearance-based gaze estimation. The architecture and configurations of the network are illustrated in Fig. 2.

During training, the inputs of our DEANet are a pair of sample-sets, $P_t$ and $P_f$. Each of them includes a left eye patch, a right eye patch, and the normalized head pose information. The components of the sample-set, acting as three streams, pass through a branch of the SNNet whose parameters are shared for both branches. In one of the Siamese branches, all patches fed into the network are a fixed-size $36 \times 60$ RGB or gray image. When the input patch is gray, it will be treated as an RGB image with the same intensity value in three channels. The normalized head pose information is a vector with a length of 2. The left eye patch and the right one are fed separately into VGG16 networks that extract the features of both patches, resulting in a vector with a length of 512. Each VGG16 network is followed by sequential operations, such as a fully connected (FC) layer with a size of 1024, a batch normalization (BN), and a rectified linear unit (ReLU) activation. The feature maps computed by each Siamese pair are concatenated (CAT), followed by another FC layer with a size of 512. After appending the normalized head pose information, other sequential operations follow, including a BN, a ReLU activation, an FC layer with a size of 256, and another ReLU activation. Lastly, the feature maps computed from both Siamese branches are concatenated, and two more FC layers with sizes of 256 and 2 follow. To avoid overfitting, a dropout layer is added before the last FC layer.

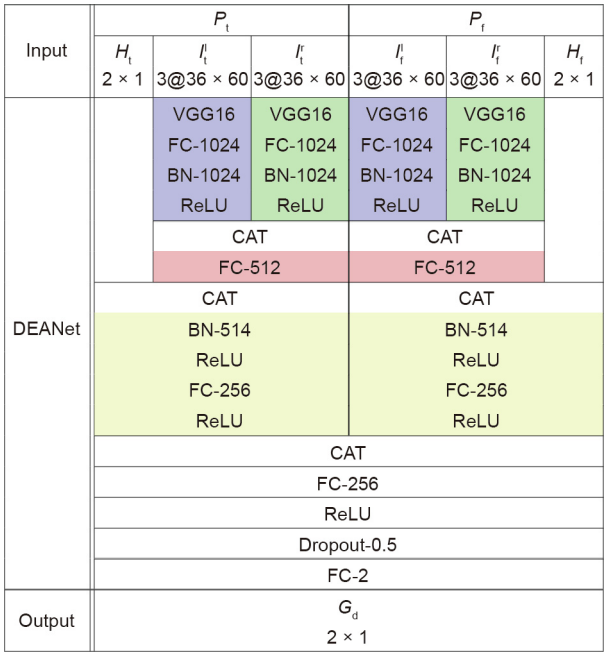


**Fig. 2.** DEANet configurations (from top to bottom). $I_t^l$, $I_t^r$, $I_f^l$, and $I_f^r$ are RGB images with a size of $36 \times 60$. $H_t$ and $H_f$ are normalized head pose information corresponding to the Siamese pairs. $G_d$ is the predicted differential gaze. All are vectors with a length of 2. VGG16 is a 16-layer Visual Geometry Group network. FC is the fully connected layer, BN is the batch normalization layer, Dropout is the Dropout layer. The layers' names are followed by their parameters. CAT is the operation that concatenates both vectors into one vector. The layers that share the same weights are highlighted by the same colors.

### 3.3.1. Siamese pair for the training phase

According to the hypothesis in this paper, a pair of labeled training samples belonging to the same person are fed into the network. Considering a dataset of $N$ training samples, there are $N^2$ possible pairs that can be used for network training. Compared with single-input algorithms [4,16,37], our proposed approach has a large number of samples for training because of the different framework. Since it is a huge value, a subset of training samples is adopted in the training phase. Strategies about the subset have been proposed in Refs. [47,48]. These are used for a classification framework where there are positive and negative pairs for both inputs. However, our proposed approach is a regressive solution that does not use explicitly positive and negative pairs. In our solution, $K < N^2$ pairs of training samples selected randomly are adopted in the training process.

### 3.3.2. Loss function

According to Fig. 1, when $G_f$ is given, the predicted $G_t$ will be close to $G_t^{gt}$ if the gaze difference predicted by DEANet is close to the ground truth differential gaze. Assume that $K$ pairs of labeled training samples $\{P_{t,k}, G_{t,k}^{gt}\}_1^K$ and $\{P_{f,k}, G_{f,k}^{gt}\}_1^K$ are given, where $G_{t,k}^{gt} \in \mathbb{R}^{2\times1}$ and $G_{f,k}^{gt} \in \mathbb{R}^{2\times1}$ are the gaze ground truth corresponding to $P_{t,k}$ and $P_{f,k}$, respectively. The loss function is formulated as follows:

$$L = \frac{1}{K}\sum_{k=0}^{K} ||G_{d,k} - G_{d,k}^{gt}||_2^2 \tag{1}$$

where the ground truth differential gaze $G_{d,k}^{gt} = G_{t,k}^{gt} - G_{f,k}^{gt}$, and $G_{d,k}$ is the differential gaze predicted by the network based on $P_{t,k}$ and $P_{f,k}$, where $||\cdot||_2$ is the $l_2$-norm operation.

### 3.4. Reference grid for the inference phase

As illustrated in Fig. 1, a gaze direction will be estimated by a labeled reference sample-set in the inference phase. The selection of the reference sample-sets will affect the estimation accuracy. Intuitively, in a good reference selection strategy, the difference between the adopted reference patches and the tested ones should not be large. A large difference will result in large errors during estimation. Moreover, a few reference sample-sets adopted in the inference phase are better than a single reference sample-set in terms of the estimation accuracy. A demonstration of the above will be discussed in Section 4.3. According to the above rules, a reference grid is then constructed in the whole gaze space, which is supported by both dimensions of the gaze directions, as shown in Fig. 3. When the difference between the input patches is small, the output of the DEANet is small as well, and vice versa. As a result, the output of the DEANet, the differential gaze, can be a metric of the distance between the reference patches and the tested ones. The evenly distributed references, as shown in Fig. 3, make the differences between some of the adopted reference patches and the tested ones so small that a promising accuracy will be achieved if the step of the grid is small enough. For example, 12 red points are the candidates for the reference gazes denoted as $G_{f,j}$, $j = 0, 1, ..., 11$. A testing gaze is marked by a blue point, which is denoted as $G_t$. Obviously, $G_t$ is computed by $G_{f,3}$, $G_{f,4}$, $G_{f,6}$, and $G_{f,7}$, rather than by other reference gazes, because the distance between $G_t$ and one of the above four reference gazes is smaller than the distance between $G_t$ and the other reference gazes. Meanwhile, because the distance between the testing gaze and the reference gaze in the gaze space can be predicted by the differential gaze in our proposed DEANet, reference gazes whose corresponding differential gazes are smaller than a certain threshold are adopted to estimate the testing gaze. To avoid empirical parameters, four reference gazes whose corresponding

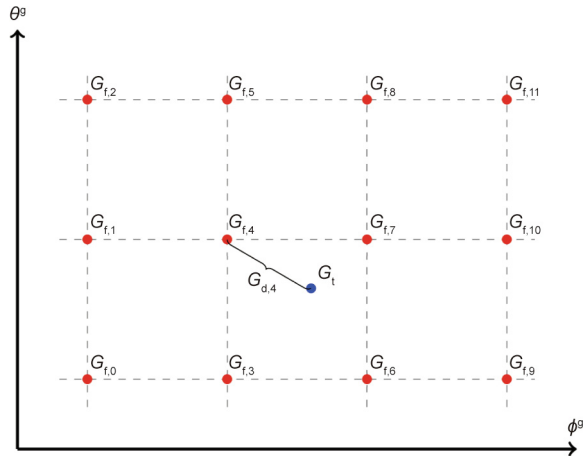


**Fig. 3.** An example of a reference grid in gaze space. Twelve reference gazes (marked with red points) are distributed in the gaze space. The blue point represents a testing gaze. The distance between $G_{f,i}$ and $G_t$ in the gaze space is predicted by the corresponding differential gaze, $G_{d,i}$.

differential gazes are smaller than the other differential gazes are adopted in this paper. After that, the testing gaze is predicted by adding each reference gaze to the corresponding differential gaze. The average value is then the final estimation. In experiments, this strategy was shown to be a good choice for all test sets.

In Ref. [25], the averaging weights are determined by comparing both feature maps extracted from the input patches. According to the construction of DEANet, the output of the network is related to the difference of both patches. Using the differential gaze as the criteria for reference selection simplifies the computation, rather than using the feature maps proposed in Ref. [25].

## 4. Experiments

### 4.1. Implementation details

Our proposed DEANet was implemented in a pytorch framework. It was trained by randomly selecting 10 000 pairs of training samples for each person. Transfer learning was utilized, and the weights of the VGG16 models were initialized by the pre-trained model [28]. An stochastic gradient descent (SGD) optimizer was adopted with a momentum of 0.9 and a weight decay of 0.0001. The batch size was 512. The initial learning rate was 0.1, and decayed by 0.1 every 5 epochs. A single GTX 1080 ti GPU was used for the network, with 20 epochs for each person.

Three experiments are reported in this section. The first experiment (Section 4.3) evaluated the DEANet based on the MPIIGaze dataset to demonstrate the reference selection strategy. The second experiment (Section 4.4) assessed the DEANet's performance in a cross-person and cross-dataset evaluation. The third experiment (Section 4.5) evaluated the DEANet against variation.

### 4.2. Datasets and protocol

The performance of the DEANet was evaluated on two public datasets, MPIIGaze and UT-Multiview. MPIIGaze was first introduced in Ref. [16]. It comprises 213 659 images from 15 participants of different ages and genders. The images were collected over different periods. To evaluate our proposed DEANet in RGB images, the eye patches and annotated gaze direction in the MPIIGaze dataset were normalized by ourselves, although some labeled gray patches and gaze directions were provided in the MPIIGaze dataset. It should be noted that the original head pose information and the target position provided by the dataset were used directly in our normalization process. UT-Multiview was initially introduced in Ref. [40]. It comprises 64 000 raw images from 50 different people. This dataset allows large amounts of synthesized eye images to be constructed by means of 3D eye shape models. UT-Multiview has a greater distribution of gaze angle than MPIIGaze. Because our introduced normalization was based on Ref. [40], the normalized patches were the same size as those in UT-Multiview. All gray patches in UT-Multiview were adopted as DEANet's training samples to evaluate the network's performance.

In experiments, a leave-one-person-out protocol was applied for the MPIIGaze dataset, while a three-fold cross-person validation protocol was used for the UT-Multiview dataset. The protocols adopted in this section are the same as other state-of-the-art algorithms [4,16,18,25,37,40].

### 4.3. Selection of reference sample-sets

In our proposed approach, the performance of the reference sample-sets will affect the estimation accuracy of the system, making the sample-sets a critical element in DEANet. In this experiment, 500 references were adopted randomly for each person in the MPIIGaze dataset. Each reference sample-set and every sample belonging to the same person made up the Siamese pairs for testing. To demonstrate the influence of the reference sample-sets on estimation accuracy, Fig. 4 illustrates the average angular error for each person in terms of references. All the Siamese pairs of each person were fed into the DEANet for gaze estimation, and the average angular error $A_t$ for each reference was formulated as follows:

$$A_t = \frac{1}{M}\sum_{m=0}^{M}\omega\left(G_{t,m}, G_{t,m}^{gt}\right) \tag{2}$$

where $M$ is the number of samples for each person in the dataset and $\omega(\cdot,\cdot)$ is the function computing the angular difference between both vectors. It should be noted that the $\omega$ function is another metric of estimation error that is equivalent to the $l_2$-norm function in Eq. (1). The $\omega$ function is intuitively adopted as the metric in experiments rather than the $l_2$-norm function for fair comparison with other algorithms adopting the same metric. As the blue bars show in Fig. 4, every person had a different estimation accuracy. Some people, such as persons No. 0, No. 1, and No. 2, had smaller angular errors than others. However, the average angular errors for other

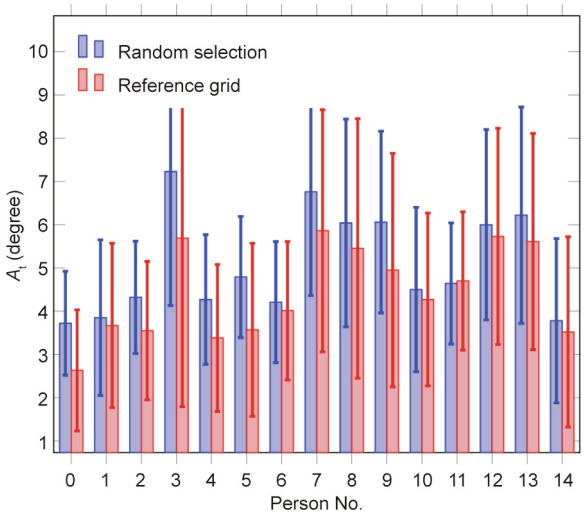


**Fig. 4.** Average angular error for each reference in the MPIIGaze dataset for different reference selection strategies: a random selection strategy, where 500 reference sample-sets were adopted randomly; and a reference grid strategy, where 12 reference sample-sets were adopted by a reference grid.

persons, such as No. 3, No. 7, No. 8, and No. 9, were much worse than those of the above persons. For example, some of the eye patches of person No. 7 included glasses, while other patches did not. If the adopted reference sample-sets did not include glasses, and the test sample-sets included glasses, their different appearances would result in large errors in the estimation accuracy, because the glasses would induce a significant amount of noise in the appearance computations. Although it is demonstrated in Ref. [16] that a generic mean facial shape model used in the normalization stage is sufficiently accurate to estimate the gaze direction, an inaccurately normalized eye patch will obviously lead to large errors in the inference stage if it is treated as a reference sample-set. Some examples are illustrated in Fig. 5.

A good reference selection strategy contributes to the improvement of the system. A key element for a reference selection strategy is to determine which patches are candidates for reference sample-sets, and which are not. This is related to the distribution of the tested samples. Fig. 6 illustrates the distribution of the 500 randomly selected reference sample-sets for persons No. 0, No. 5, and No. 7 in the gaze space. Each reference gaze can be represented by a point in the gaze space. When the average angular error of reference $i$, $A_{t,i}$, is smaller than the mean of all the references, the corresponding reference is identified as a "good" reference (marked in red in Fig. 6). Conversely, when $A_{t,i}$ is greater than the mean of all the references, the corresponding reference is identified as a "bad" reference (marked in blue in Fig. 6). The gray points are all the samples that are used to represent the whole distribution for each person. In Fig. 6, bad references are almost all located at the periphery of the whole distribution, especially in person No. 7, while good references are evenly distributed in the whole space. Some sample-sets that include large gaze directions cannot be selected as reference sample-sets. Moreover, a single reference strategy is not sufficient for accurate estimation.

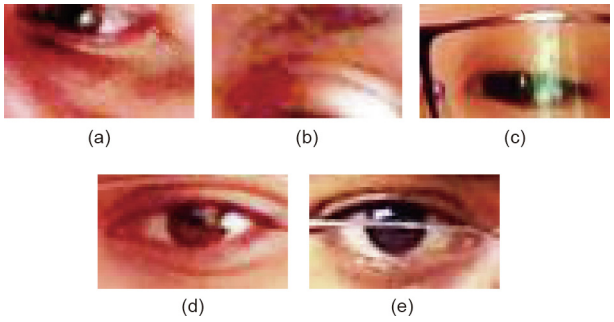


**Fig. 5.** Examples of normalized patches that result in large errors. (a, b) Inaccurately normalized eye patches (p03-day54-0097-left and p08-day31-0301-left). (c) Noise induced by glasses (p09-day12-0158-left). (d, e) An image without glasses as the reference sample-set (p07-day24-0046-left) and an image with glasses as the test one (p07-day25-0255-right). The name of each patch comes from the MPIIGaze dataset.

Fig. 6 suggests that the distribution of the reference sample-sets affects the system's performance. Furthermore, the difference between the reference and tested sample-sets also has an influence on the system's performance. It should be noted that the ground truth difference between both sample-sets can be represented by the ground truth angular error between $G_t^{gt}$ and $G_f^{gt}$ according to our proposed network. Moreover, the system's estimation error can be formulated as $\omega(G_t, G_f^{gt})$. This is also the predicted value of the ground truth difference between both sample-sets. The relationship between the difference of both sample-sets and the estimation accuracy is illustrated in Fig. 7. In order to simplify the figure, $\omega(G_t^{gt}, G_f^{gt})$ was quantified into 100 bins, and $\omega(G_t, G_f^{gt})$ was accordingly the mean value. These are denoted as $\overline{\omega}(G_t^{gt}, G_f^{gt})$ and $\overline{\omega}(G_t, G_f^{gt})$ in Fig. 7, respectively. The estimation error will increase when the difference between the testing gaze and the reference gaze increases. A good reference gaze direction close to the testing gaze direction will obtain a good estimation accuracy. Because the testing gaze directions are not provided, more reference sample-sets will be involved. This is a trade-off between the number of references and the estimation accuracy. Moreover, although the testing gaze directions are not provided, the scales of the testing gaze should be known in advance. The reference grid can be constructed according to the scale of the gaze directions. In our proposed approach, a three-row and four-column grid was constructed in order to obtain a good performance in all experiments. Examples are illustrated in Fig. 6 with green points. Accordingly, the DEANet with the reference grid was evaluated using the MPIIGaze dataset for each person; the average angular errors are illustrated in Fig. 4 (red bars). The results suggest that almost all the average angular errors with a reference grid strategy are better than the errors with a random selection strategy. The mean angular error for all the persons decreases from 5.09 for the random selection strategy to 4.38 for the reference grid strategy, so a 14% improvement in performance is achieved using the reference grid strategy.

### 4.4. Cross-person and cross-dataset evaluations

The proposed DEANet is a person-independent model that can estimate the gaze direction for a new person. Information for the new person is incorporated into the network as reference sample-sets in the testing stage. Thus, the problem of a person-independent model being irrelevant to a new person is effectively avoided. In order to evaluate how well the DEANet addresses the challenge, a cross-person evaluation was performed in both public datasets. Table 1 illustrates the mean angular errors of the proposed algorithm and of other approaches based on the

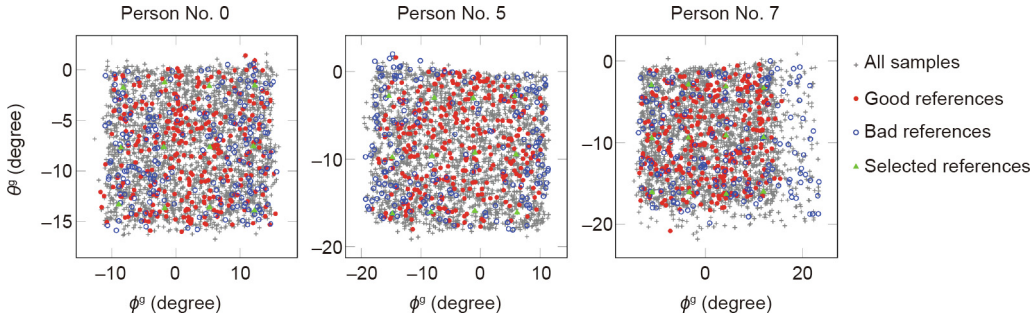


**Fig. 6.** Distributions of gaze angle for persons No. 0, No. 5, and No. 7 in MPIIGaze. Any reference sample-set can be represented by a point in the gaze angle dimension in terms of its labeled gaze direction. The red points are good reference sample-sets whose value, $A_t$, is smaller than the mean value of all the references; blue points are bad reference sample-sets whose value is greater than the mean value of all the references. Gray points are all the samples for each person. Green points are the adopted references according to the reference grid in our experiments.

MPIIGaze and UT-Multiview datasets. Our proposed algorithm achieves favorable results in both datasets. Although the same SNNet framework was adopted by both Ref. [25] and our proposed approach, the performance of our proposed approach is better than that in Ref. [25] because our approach involves more information, including the information for both eyes and the head pose. Compared with MPIIGaze, the UT-Multiview dataset includes more people, so the performance of all algorithms evaluated on UT500 Multiview are better than those evaluated on MPIIGaze. As data-driven models, diversity of the training data increases the performance of the pre-trained models, and our proposed DEANet outperforms the other algorithms in both datasets.

To demonstrate the robustness of our proposed approach, a cross-dataset evaluation was performed as well. The model was trained on the UT-Multiview dataset and then tested on the MPIIGaze dataset. Fig. 8 illustrates the mean angular errors of all the evaluated algorithms for the cross-dataset evaluation [16,29,40,51,52]. Because the gaze distribution of the training samples is different from the distribution of the testing ones, all algorithms performed worse in the cross-dataset evaluations than in the cross-person evaluations. However, our proposed DEANet is a differential network, and the input and output of the network are replaced with differential inputs and outputs. Our proposed approach is more robust against gaze distributions than other traditional methods. The mean angular error of our proposed approach is 7.77 degrees, with a standard deviation of 3.5 degrees.

### 4.5. Performance against variation

In the previous evaluations, our proposed DEANet achieved a good performance in gaze estimation. In this section, the performance against variation, such as the influence of the head pose information and the image resolution, is further investigated. To deal with arbitrary head pose information in our proposed DEANet, a normalized head pose information was adopted. To demonstrate the performance of the DEANet against variation, a cross-person evaluation in the MPIIGaze dataset without the head pose information was performed. In this experiment, a new network without the head pose information was retrained based on the MPIIGaze dataset. The mean angular error evaluated for all persons was 4.46, which is a little higher than that for the network with the head pose information (4.38), as reported in Table 1. The network's performance will be slightly degraded without the head pose information. The head pose information is marginal for a deep network such as DEANet. However, it is still important for a shallower network, such as MnistNet [53], which is evaluated in Ref. [16]. Shallower networks are usually adopted in order to save computation resources, especially in remote devices.

Moreover, the influence of the image resolution on gaze estimation was investigated in this experiment. The same network parameters were adopted as those proposed in Section 4.4, and the cross-person evaluation was performed. The protocols were the same as those described in Section 4.4. In the evaluation, all the input patches were resized to $18 \times 30$, $9 \times 15$, and $5 \times 8$. It should be noted that the resized patches needed to be restored to the original size ($36 \times 60$) by interpolation in order to be successfully fed into the DEANet. The DEANet's performance for a different image resolution was compared with that of GazeNet [16] based on both the MPIIGaze and UT-Multiview datasets, as shown in Table 2. Our proposed DEANet outperforms GazeNet in this experiment.

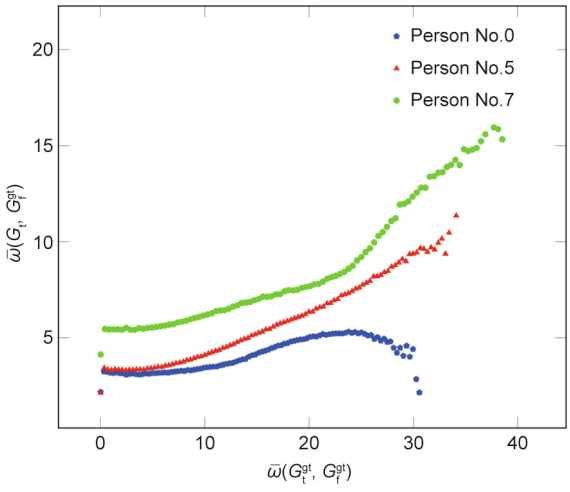


**Fig. 7.** The relationship between the estimation error ($y$-axis) and the difference between both sample-sets ($x$-axis) for persons No. 0, No. 5, and No. 7.

**Table 1**
Gaze directional results on two popular datasets with mean angular error (in degrees).

| Method | MPIIGaze | UT-Multiview |
|---|---|---|
| GazeNet [16] | 5.5 | 4.4 |
| Diff-NN [25] | 4.64 | 4.13 |
| RT-GENE Net [18] | 4.8 | – |
| iTracker [4] | 5.6 | – |
| Full face [37] | 4.8 | – |
| MnistNet [29] | 6.1 | – |
| LbS [40] | 6.7 | 6.5 |
| Ours | 4.38 | 3.56 |

A leave-one-person-out protocol was used in the MPIIGaze dataset, and a three-fold

## 5. Conclusions

This paper presented a novel DEANet for appearance-based gaze estimation. Three streams—including both eye patches and

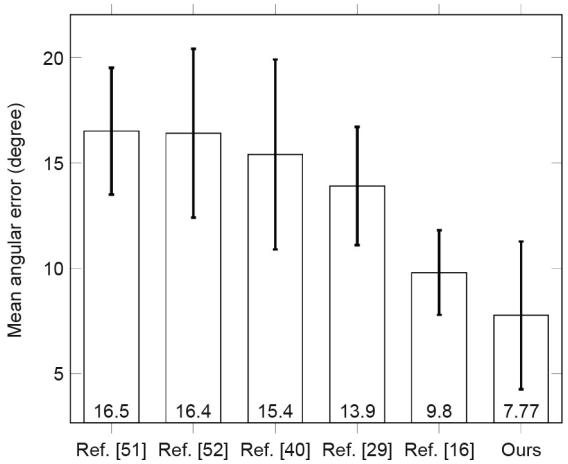


**Fig. 8.** Mean angular error for a cross-dataset evaluation with training on the UT-Multiview dataset and testing on the MPIIGaze dataset.

**Table 2**
The influence of image resolution. Mean angular errors were evaluated on the MPIIGaze and UT-Multiview datasets with different image resolutions.

| Image resolution | MPIIGaze | | UT-Multiview | |
|---|---|---|---|---|
| | Ours | GazeNet [16] | Ours | GazeNet [16] |
| $18 \times 30$ | 5.41 | – | 3.75 | 9.9 |
| $9 \times 15$ | 8.57 | – | 5.42 | 11.4 |
| $5 \times 8$ | 12.10 | – | 13.07 | 15.7 |
| Average | 8.69 | 11.7 | 7.41 | 12.3 |

the head pose information—are fed into the network, and a person-independent model is trained based on an SNNet framework. Because the differential gaze is adopted, person-specific information can be used in the testing stage. A reference grid is constructed for reference candidates, and the proposed strategy selects good references to improve the estimation accuracy. Our approach was evaluated on two public datasets: MPIIGaze and UT-Multiview. The extensive experimental evaluations showed that our approach achieves a more promising performance than other popular methods.

All experiments were analyzed theoretically on the public datasets. Our proposed approach will be encompassed as a modality for HRC robot control with multimodal fusion, which will be investigated carefully in our future work.

## Acknowledgements

This work was supported by the Science and Technology Support Project of Sichuan Science and Technology Department (2018SZ0357) and China Scholarship.

## Compliance with ethics guidelines

Song Gu, Lihui Wang, Long He, Xianding He, and Jian Wang declare that they have no conflict of interest or financial conflicts to disclose.

## References

[1] Palinko O, Rea F, Sandini G, Sciutti A. Robot reading human gaze: why eye tracking is better than head tracking for human-robot collaboration. In: Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2016 Oct 9–14; Daejeon, Republic of Korea. New York: IEEE; 2016. p. 5048–54.
[2] Duarte NF, Rakovic M, Tasevski J, Coco MI, Billard A, Santos-Victor J. Action anticipation: reading the intentions of humans and robots. IEEE Robot Autom Lett 2018;3(4):4132–9.
[3] Thies J, Zollhöfer M, Stamminger M, Theobalt C, Niener M. FaceVR: real-time facial reenactment and eye gaze control in virtual reality. ACM T Graphic 2018;37(2):1–15.
[4] Krafka K, Khosla A, Kellnhofer P, Kannan H, Bhandarkar S, Matusik W, et al. Eye tracking for everyone. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. New York: IEEE; 2016. p. 2176–84.
[5] Liu H, Wang L. Gesture recognition for human-robot collaboration: a review. Int J Ind Ergon 2018;68:355–67.
[6] Liu H, Wang L. Human motion prediction for human-robot collaboration. J Manuf Syst 2017;44(Pt 2):287–94.
[7] Wang L. From intelligence science to intelligent manufacturing. Engineering 2019;5(4):615–8.
[8] Liu H, Fang T, Zhou T, Wang L. Towards robust human–robot collaborative manufacturing: multimodal fusion. IEEE Access 2018;6:74762–71.
[9] Day CP. Robotics in industry-their role in intelligent manufacturing. Engineering 2018;4(4):440–5.
[10] Bulling A, Roggen D, Tröster G, Tröster G. Wearable EOG goggles: seamless sensing and context-awareness in everyday environments. J Ambient Intell Smart Environ 2009;1(2):157–71.
[11] Hansen DW, Ji Q. In the eye of the beholder: a survey of models for eyes and gaze. IEEE Trans Pattern Anal Mach Intell 2010;32(3):478–500.
[12] Valenti R, Sebe N, Gevers T. Combining head pose and eye location information for gaze estimation. IEEE Trans Image Process 2011;21(2):802–15.
[13] Alberto Funes Mora K, Odobez JM. Geometric generative gaze estimation (G3E) for remote RGB-D cameras. In: Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. New York: IEEE; 2014. p. 1773–80.
[14] Zhang X, Sugano Y, Bulling A. Evaluation of appearance-based methods and implications for gaze-based applications. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; 2019 May; Glasgow Scotland, UK. New York: Association for Computing Machinery; 2019. p. 1–13.
[15] Lu W, Li Y, Cheng Y, Meng D, Liang B, Zhou P. Early fault detection approach with deep architectures. IEEE Trans Instrum Meas 2018;67(7):1679–89.
[16] Zhang X, Sugano Y, Fritz M, Bulling A. MPIIGaze: real-world dataset and deep appearance-based gaze estimation. IEEE Trans Pattern Anal Mach Intell 2019;41(1):162–75.
[17] Yu Y, Liu G, Odobez JM. Deep multitask gaze estimation with a constrained landmark-gaze model. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 9–14; Munich, Germany. New York: Springer; 2018. p. 456–74.
[18] Fischer T, Jin Chang H, Demiris Y. Rt-gene: real-time eye gaze estimation in natural environments. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 9–14; Munich, Germany. New York: Springer; 2018. p. 334–52.
[19] Choe KW, Blake R, Lee SH. Pupil size dynamics during fixation impact the accuracy and precision of video-based gaze estimation. Vision Res 2016;118:48–59.
[20] Guestrin ED, Eizenman M. General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Trans Biomed Eng 2016;53(6):1124–33.
[21] Lu F, Sugano Y, Okabe T, Sato Y. Adaptive linear regression for appearance-based gaze estimation. IEEE Trans Pattern Anal Mach Intell 2014;36(10):2033–46.
[22] Huang MX, Kwok TC, Ngai G, Leong HV, Chan SC. Building a self-learning eye gaze model from user interaction data. In: Proceedings of the 22nd ACM international conference on Multimedia; 2014 Nov; Orlando Florida, USA. New York: Association for Computing Machinery; 2014. p. 1017–20.
[23] Liu G, Yu Y, Mora KAF, Odobez JM. A differential approach for gaze estimation. IEEE Trans Pattern Anal Mach Intell 2019;43(3):1092–9.
[24] Venturelli M, Borghi G, Vezzani R, Cucchiara R. From depth data to head pose estimation: a siamese approach. In: Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (Volume 5); 2017 Feb 27–Mar 1; Porto, Portugal. New York: Springer; 2017. p. 194–201.
[25] Liu G, Yu Y, Mora KAF, Odobez JM. A differential approach for gaze estimation with calibration. IEEE Trans Pattern Anal Mach Intell 2021;43(3):1092–9.
[26] Bromley J, Guyon I, Lecun Y, Säckinger E, Shah R. Signature verification using a "siamese" time delay neural network. In: Proceedings of the 6th International Conference on Neural Information Processing Systems; 1993 Nov; Denver, CO, USA: Morgan Kaufmann Publishers; 1994. p. 737–44.
[27] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 1; 2014 Dec. Cambridge: MIT Press; 2014. p. 568–76.
[28] Simonyan K, Zisserman A. Very deep convolutional networks for largescale image recognition. In: Proceeding of International Conference on Learning Representations 2015; 2015 May 7–9; San Diego, CA, USA. New York: WikiCFP; 2015.
[29] Zhang X, Sugano Y, Fritz M, Bulling A. Appearance-based gaze estimation in the wild. In: Proceeding of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. New York: IEEE; 2015. p. 4511–20.
[30] Lian D, Hu L, Luo W, Xu Y, Duan L, Yu J, et al. Multiview multitask gaze estimation with deep convolutional neural networks. IEEE Trans Neural Netw Learn Syst 2019;30(10):3010–23.
[31] Park S, Spurr A, Hilliges O. Deep pictorial gaze estimation. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 9–14; Munich, Germany. New York: Springer; 2018. p. 741–57.
[32] Liu J, Francis BSL, Rajan D. Free-head appearance-based eye gaze estimation on mobile devices. In: Proceeding of 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC); 2019 Feb 11–13; Okinawa, Japan. New York: IEEE; 2019. p. 232–7.
[33] Wong ET, Yean S, Hu Q, Lee BS, Liu J, Deepu R. Gaze estimation using residual neural network. In: Proceeding of 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops); 2019 Mar 11–15; Kyoto, Japan. New York: IEEE; 2019. p. 411–4.
[34] Dubey N, Ghosh S, Dhall A. Unsupervised learning of eye gaze representation from the web. In: Proceeding of 2019 International Joint Conference on Neural Networks (IJCNN); 2019 Jul 14–19; Budapest, Hungary. New York: IEEE; 2019. arXiv:1904.02459v1.
[35] Funes-Mora KA, Odobez JM. Gaze estimation in the 3D space using RGB-D sensors. Int J Comput Vis 2016;118(2):194–216.
[36] Funes Mora KA, Monay F, Odobez JM. Eyediap: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras. In: Proceeding of the Symposium on Eye Tracking Research and Applications; 2014 Mar 26–28; Florida, UF, USA. New York: Association for Computing Machinery; 2014. p. 255–8.
[37] Sugano Y, Fritz M, Andreas Bulling X, et al. It's written all over your face: full-face appearance-based gaze estimation. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2017 Jul 21–26; Honolulu, HI, USA. New York: IEEE; 2017. p. 51–60.
[38] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proceeding of the 25th International Conference on Neural Information Processing Systems—Volume 1; 2012 Dec 14–16; Siem Reap, Cambodia. LaneRed Hook: Curran Associates Inc; 2012. p. 1097–105.
[39] Ogusu R, Yamanaka T. Lpm: learnable pooling module for efficient full-face gaze estimation. In: Proceeding of 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019); 2019 May 14–18; Lille, France. New York: IEEE; 2019. p. 1–5.
[40] Sugano Y, Matsushita Y, Sato Y. Learning-by-synthesis for appearance-based 3D gaze estimation. In: Proceeding of 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. New York: IEEE; 2014. p. 1821–8.

[41] Zhang X, Sugano Y, Bulling A. Revisiting data normalization for appearance-based gaze estimation. In: Proceeding of the 2018 ACM Symposium on Eye Tracking Research & Applications. 2018 Jun 14–17; Warsaw, Poland. New York: Association for Computing Machinery; 2018. p. 1–9.
[42] Sugano Y, Matsushita Y, Sato Y, Koike H. An incremental learning method for unconstrained gaze estimation. In: Proceeding of European Conference on Computer Vision; 2008 Oct 12–18; Marseille, France. Berlin: Springer; 2008. p. 656–67.
[43] Zhang X, Huang MX, Sugano Y, Bulling A. Training person-specific gaze estimators from user interactions with multiple devices. In: Proceeding of the 2018 CHI Conference on Human Factors in Computing Systems; 2018 Apr 21–26; Montréal, QC, Canada. New York: Association for Computing Machinery; 2018. p. 624.
[44] Yu Y, Liu G, Odobez JM. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In: Proceeding of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. New York: IEEE; 2019. p. 2019.
[45] Veges M, Varga V, Lőrincz A, András L. 3D human pose estimation with Siamese equivariant embedding. Neurocomputing 2019;339:194–201.
[46] Doumanoglou A, Balntas V, Kouskouridas R, Kim TK. Siamese regression networks with efficient mid-level feature extraction for 3D object pose estimation. In: Proceeding of 29th Conference on Neural Information Processing Systems (NIPS 2016); 2016 Dec 5–10; Barcelona, Spain; 2016.
[47] Simo-Serra E, Trulls E, Ferraz L, Kokkinos I, Fua P, Moreno-Noguer F. Discriminative learning of deep convolutional feature point descriptors. In: Proceeding of 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. New York: IEEE; 2015. p. 118–26.
[48] Wang J, Song Y, Leung T, Rosenberg C, Wang J, Philbin J, et al. Learning fine-grained image similarity with deep ranking. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. Washington, DC: IEEE Computer Society; 2014. p. 1386–93.
[49] Baltrusaitis T, Robinson P, Morency LP. Continuous conditional neural fields for structured regression. In: Proceeding of European conference on computer vision; 2014 Sep 6–12; Zurich, Switzerland. Berlin: Springer; 2014. p. 593–608.
[50] Lepetit V, Moreno-Noguer F, Fua P. EP*n*P: an accurate o(n) solution to the P*n*P problem. Int J Comput Vis 2009;81(2):155–66.
[51] Schneider T, Schauerte B, Stiefelhagen R. Manifold alignment for person independent appearance-based gaze estimation. In: Proceeding of 2014 22nd International Conference on Pattern Recognition; 2014 Aug 24–28; Stockholm, Sweden. New York: IEEE; 2014. p. 1167–72.
[52] Mora KAF, Odobez JM. Person independent 3D gaze estimation from remote RGB-D cameras. In: Proceeding of 2013 IEEE International Conference on Image Processing. 2013 Sep 15–18; Melbourne, Australia. New York: IEEE; 2013. p. 2787–91.
[53] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–324.