Deep Learning-based Pupil Center Detection for Fast and Accurate Eye Tracking System

Kang Il Lee, Jung Ho Jeon and Byung Cheol Song

Department of Electrical Engineering, Inha University, Incheon, Korea
{kangil-lee, jungho-jeon}@inha.edu, bcsong@inha.ac.kr

Abstract. In augmented reality (AR) or virtual reality (VR) systems, eye tracking is a key technology and requires significant accuracy as well as real-time operation. Many techniques for detecting pupil centers with error range of iris radius have been developed, but few techniques have precise performance with error range of pupil radius. In addition, the conventional methods rarely guarantee real-time pupil center detection in a general-purpose computer environment due to high complexity. Thus, we propose more accurate pupil center detection by improving the representation quality of the network in charge of pupil center detection. This is realized by representation learning based on mutual information. Also, the latency of the entire system is greatly reduced by using non-local block and self-attention block with large receptive field, which makes it accomplish real-time operation. The proposed system not only shows real-time performance of 52 FPS in a general-purpose computer environment but also provides state-of-the-art accuracy in terms of fine level index of 96.71%, 99.84% and 96.38% for BioID, GI4E and Talking Face Video datasets, respectively.

Keywords: Remote Eye Tracking, Mobile Applications

1 Introduction

In general, the performance of computer vision applications such as AR and VR highly depend on gaze estimation and eye tracking techniques. The pupil center detection plays a crucial role in those applications. As a result, the real-time operation and high accuracy of pupil center detection make the AR/VR system more practical. Pupil center detection or tracking (PCT) methods are divided into two categories, i.e., model-based and appearance-based approaches. Model-based methods are limitedly used in equipments such as head-mounted goggles. For example, user-specific calibrations were performed using eye geometry models and coordinate systems for accurate pupil center detection at a close distance [1].

On the other hand, appearance-based methods [2–12] detect the pupil center using a remote camera without head-mount goggles or user-specific calibrations.In general, appearance-based methods consist of two steps: the eye region extraction and the pupil center detection. The eye region extraction step is again

composed of a face detection and an eye region extraction using features such as landmarks. The pupil center detection methods can be prior knowledge-based approach or context-based approach. Prior knowledge-based approaches adopt a regression model designed based on generic eye appearance information [2, 4–11]. The context-based techniques separate a given eye region image into a pupil center and a background using specific segmentation networks [12, 13]. Alternatively, [3] improved the performance of the regression model with the hand-crafted features extracted from the pupil area.

Most of the latest PCT methods provide high accuracy in the error range of iris radius, whereas their accuracy in the error range of pupil radius, a more precise level, is not satisfactory yet. In addition, many previous techniques seldom guarantee real-time PCT in a general purpose computer environment due to high latency of eye region extraction modules [4, 10, 11].

This paper proposes a new appearance-based PCT to secure real-time operation as well as high accuracy even at the precise level. We propose the PCT system to enable pupil center detection robust against glasses wearing, inspired by [13]. First, representation learning using mutual information (MI) is applied to the pupil center detection network so that the network can extract features which have rich location information. Note that as the representation of the network improves, the pupil center detection accuracy increases together. Next, in order to realize consistent real-time processing of the PCT system, nonlocal block (NLB) and self-attention block (SAB) are applied to face detection network and glasses removal network, which are bottlenecks in terms of latency. Using NLB and SAB, each network can obtain a low latency because large receptive field effect is produced even with only a few layers. On the other hand, glasses removal network tends to blur the eye region during erasing glasses. Therefore, we propose a method to mitigate blur phenomenon by employing perceptual loss, resulting in improvement of detection performance.

The main contribution points of the proposed PCT system are as follows.

- Overall pupil detection accuracy was increased by improving representation quality of pupil detection network through representation learning using MI.
- The latency of the entire system was greatly reduced by the face detection network and the glasses removal network which utilize the large receptive field features of NLB and SAB. That is, it guarantees real-time operation.
- The spatial loss or blur due to the structural lightweighting of the glasses removal network is compensated by employing perceptual loss.
- In terms of fine-level index [4], i.e., most precise accuracy level, the proposed PCT system shows state-of-the-art(SOTA) performance of 96.71%, 99.84%, and 96.38% for BioID, GI4E, and Talking Face Video datasets, respectively.

2 Related Works

Appearance-based Methods. [2], which is a representative appearance-based method using prior knowledge, first detects a face from an input image through the face detector of [14]. The eye area is then cropped using biometric statistics

from the face. Next, the radial gradients for all the points on the iris contour are computed. Finally, a pupil center is detected by using the prior knowledge that the origin of the displacement vector with the maximum radial gradient is the pupil center. One of the latest techniques using prior knowledge is based on cascade regression model and circle fitting [4]. This algorithm trains three cascade regression models: two regression models for eye corners and eye centers, and a regression model for circle fitting.

On the other hand, appearance-based techniques considering the context have been reported [12, 13]. For instance, [13] detects a face using [15], and then determines whether there is glasses in the face. If glasses exist, an eye region is detected after removing the glasses using CycleGAN [16]. Otherwise, the eye region detection module is activated immediately. Finally, a pupil center is detected by applying the semantic segmentation network to the detected eye region so as to separate the pupil center from the background. [12] detects the eye center by directly inputting the face image to the semantic segmentation network. Note that [12] does not include any separate module for dealing with glasses.

Non-local Neural Networks. The most popular convolutional neural network (CNN) is based on convolutional operations for local neighborhood. Therefore, a deep network is required to secure a large receptive field. However, this approach causes a gradient vanishing problem, making learning difficult and inefficient. In order to solve this fundamental limitation, NLB was born in [17]. If NLB receives an embedded feature map, it measures point-to-point graphical relations for all points in the feature map. The measured graphical relation can produce a self-attention effect similar to [18]. This helps to solve a given task efficiently. In addition, since the receptive field size of NLB is the same as that of the input data, NLB generally has an effect of enlarging a receptive field. Therefore, if NLB is applied to a network, the network can achieve a large receptive field effect without stacking deep convolutional layers.

Thus, we apply NLB to the face detection network that is a bottleneck in terms of latency so that overall latency can be decreased. In addition, SAB, which has a structure similar to NLB, is applied to another bottleneck, i.e., the glasses removal network, to maximize operational speed.

Perceptual Loss. In typical image processing tasks such as noise reduction, super-resolution, and colorization, traditional pixel loss does not reflect perceptual characteristics well. With this in mind, [19] defines a loss in feature level, i.e., perceptual loss. Then high level features are extracted by VGG-16 [20] that is pre-trained with ImageNet [21]. The extracted features are used to induce perceptual loss. Inspired by the image super resolution technique proposed in [19], we add perceptual loss to the cycle consistency loss [16] defined for learning the lightweight glasses removal network.

Mutual Information Maximization. In information theory, the mutual information (MI) of two random variables indicates the mutual dependence between the two random variables. More specifically, MI quantifies the amount of information of one random variable through observing the other random variable. Representation learning aimed at maximizing MI between target representation

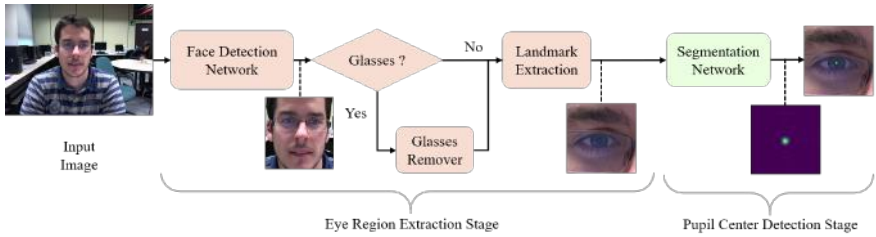

Fig. 1. The proposed pupil center tracking system.

and features or intermediate features has been studied for a long time [22–24]. Most recent studies focus on variational approach [24, 25]. This is because it is very difficult to estimate the MI of continuous random variables in high dimensional space. So the variational approach derives the tight lower bound of MI. Recently, a few methods have proposed tight lower bounds of MI estimation by using neural network [22, 23]. For example, [22] achieved high performance in downstream tasks through representation learning using MI maximization. Inspired by [22], we apply MI-based representation learning to the segmentation network. This greatly improves the pupil center detection accuracy.

3 Method

Fig. 1 is the overview of the proposed PCT system (Sec.3.1). We designed the entire system, inspired by [13], which combines glasses removal module with the structure of a universal pupil center detection scheme. Even though we follow the basic structure of [13], we propose a few novel methods for dramatically improving the speed and accuracy of the PCT system. First, we present a low latency face detection network using NLB (Sec. 3.2). Next, we propose a structure applying SAB for the low latency of the glasses removal module that is the critical latency bottleneck of the entire PCT system in [13]. Plus, perceptual loss is introduced to mitigate the blur phenomenon (Sec. 3.3). Finally, we propose a representation learning using MI maximization to improve the representation quality of the segmentation network which has an absolute influence on the overall accuracy of the PCT system (Sec. 3.4).

3.1 Overview

This section describes the purpose and operation of each module of the proposed PCT system. As shown in Fig. 1, the proposed PCT system consists of an eye region extraction stage and a pupil center detection stage. The face detection network, the first step of the eye region extraction stage, is responsible for cropping the face from an input image. Next, the glasses classifier determines whether glasses exist in the face image. If it is determined that the glasses are worn, the glasses removal network removes the glasses. Next, the landmarks are extracted

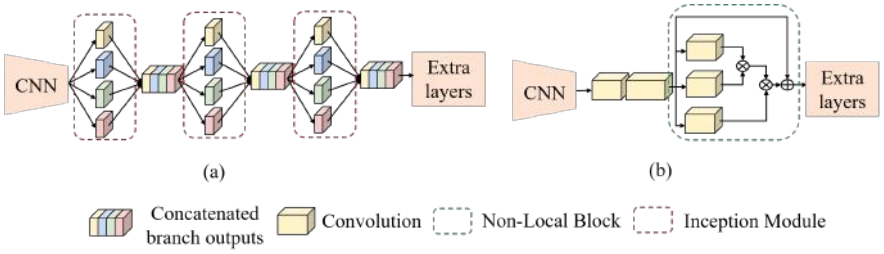

Fig. 2. (a) FaceBoxes [26] (b) the modified face detection network. CNN is the encoder network and extra layers are convolution layers designed to address multi-scale in [26]

from the face image, while glasses are removed selectively. The final step of the eye region extraction stage is to extract the eye region in the face image using landmark information. Subsequent pupil center detection stage is performed by a segmentation network. Given an eye region image, the heat map is calculated as shown in the lower right corner of Fig. 1. The glasses classifier and landmark extractor of the proposed PCT system follow the methods of [13].

3.2 Face Detection Network

The face detection networks used in conventional PCT methods can have a high accuracy, but requires a long latency to guarantee high precision. So we choose a high-performance face detection network[26], and apply NLB to the network to reduce the latency while maintaining reasonable accuracy.

Faceboxes of [26] effectively used context information through an inception module[27] while properly coping with multi-scale like SSD [15]. So [26] is robust for scale and occlusion. However, since the inception module is basically a concatenation structure, it can be a bottleneck when operating on GPU. To remedy this problem, we introduce NLB that can apply context information, which secures a large receptive field with only a few layers. In detail, the inception module is replaced with the NLB having two convolutional layers, as shown in Fig. 2(b). As a result, the modified FaceBoxes network provides high face detection performance with low latency (see the results in Section 4).

3.3 Glasses Removal Network

This section describes how to lighten the glasses removal network, which was proposed in [13]. Also, how to mitigate the blur phenomenon during light-weighting is given.

The CycleGAN[16] was used to refine landmark information as the glasses removal network[13]. But the CycleGAN generator used in [13] has a structural problem. Since the encoder has only two down-sampling layers as shown in Fig. 3 (a), it faces with a critical problem that the computational cost varies depending on the resolution of an input image. In addition, constructing a network by

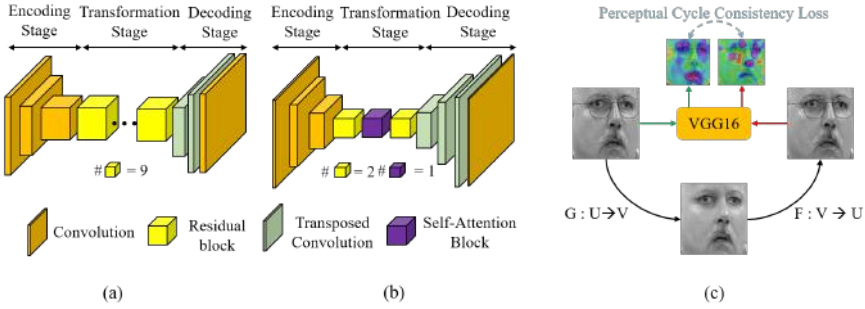

Fig. 3. (a) The glasses removal network in [13] (b) the proposed glasses removal network (c) the illustration of how to calculate perceptual cycle consistency loss.

just stacking nine residual blocks in a high level layer is an inefficient configuration[18]. To address the computational cost problem caused by the above-mentioned structural factor of the existing glasses removal network, we attempt the following approach. First, add a down-sampling layer and an up-sampling layer to the encoder and decoder as shown in Fig. 3(b) to reduce the spatial size of the feature map for the transformation stage to 1/4. Also, the transformation stage is configured to have a comparable receptive field to the generator of CycleGAN by using one self-attention block [28] and two residual blocks.

However, this structural approach causes additional spatial loss in encoding process. So, we address the spatial loss problem by utilizing the method proposed in [19]. Since CycleGAN basically learns using unpaired datasets, there is no high resolution (HR) reference for the domain translated image. Thus, we propose the cycle consistency loss in feature-level as shown in Fig. 3(c). We call this perceptual cycle consistency loss (PCCL), which is defined as Eq. (1).

$$L_{PCCL} = E_{u\sim P_{data(U)}}\left[\sum_{i=1}^{5}\left|\phi_i(F(G(u))) - \phi_i(u)\right|_1\right] + E_{v\sim P_{data(V)}}\left[\sum_{i=1}^{5}\left|\phi_i(G(F(v))) - \phi_i(v)\right|_1\right] \tag{1}$$

where U and V mean different domains. G and F denote a generator for removing the glasses and a generator for restoring glasses, respectively. In addition, $\phi_i(\cdot)$ stands for the feature map of the last layer per unit of five convolution blocks of VGG-16[20]. By integrating GAN loss L_{GAN} of [16] and L_{PCCL}, total GAN loss L_{GAN}^{Total} is defined as Eq. (2).

$$L_{GAN}^{Total} = L_{GAN} + L_{PCCL} \tag{2}$$

As a result, we construct the glasses removal network with significantly lower computational cost than [13] through the proposed light-weighting strategy. Since PCCL compensates for the spatial loss due to the light-weighting, we can keep the performance comparable to the glasses removal network of [13].

3.4 Segmentation Network

This section proposes a method to improve the accuracy of the proposed PCT system. We enhance the representation quality of the pupil center detection network, i.e., the segmentation network, through representation learning using MI maximization. [13] showed that low-level feature transfer to decoder through skip connection can improve pupil center detection performance. Also, the pupil center detection performance was further improved by enhancing the representation of latent features through an auxiliary network of auto-encoder structure. Therefore, we found that the representation quality of the pupil center detection network greatly influences the overall pupil center detection performance.

On the other hand, in [22], the MI between the local feature and the representation of the auto-encoder was measured, and then unsupervised representation learning was conducted by maximizing the MI. [22] showed that the learned encoder provides high performance in classification, that is a kind of downstream task such as semantic segmentation task.

Based on the results of [13] and [22], in order to maximize the pupil center detection performance of the segmentation network, we propose a method of enhancing the representation quality through MI maximization only during training. First, the low-level feature X_1 and latent feature X_2 of the segmentation network and the feature map Y of the decoder are extracted, as shown in Fig. 4. We define X_1 and X_2 as local features and Y as the representation of segmentation network. Then, to calculate the MI regardless of the spatial dimensions of X_1 and X_2, we transform Y into a feature vector $\tilde{Y}$ through the vectorization network. $\tilde{Y}$ is input to the shuffle and concatenation module together with X_1 and X_2. Next, the concatenated feature maps are produced for computing the conditional entropy (CE) and marginal entropy (ME) estimates. If the CE and ME estimates enter two discriminators D_1 and D_2, the MI is calculated according to Eq. (3).

$$I(X_k;\tilde{Y}) = E_{\mathbb{P}\otimes\mathbb{P}'}[sp(D_k(C(X_k^{'})))] - E_{\mathbb{P}}[-sp(D_k(C(X_k,\tilde{Y})))] \tag{3}$$

where sp indicates the softplus operator and C means the concatenation operation. And $\mathbb{P}$ is the distribution of an input local feature X_k. $X_k^{'}$ is the local feature processed by batch-wise shuffle as in Fig. 4, and $\mathbb{P}^{'}$ is the distribution of $X_k^{'}$ (k=1,2). Finally, by adding MI to the segmentation loss L_{Seg} [13] based on pixel-wise mean squared error, the total loss function L_{total} of the segmentation network is defined by Eq. (4).

$$L_{total} = L_{seg} - (I_1(X_1;\tilde{Y}) + I_2(X_2;\tilde{Y})) \tag{4}$$

The operation of the proposed segmentation network is summarized as follows:

- Define local features(X_1, X_2) and representation(Y) suitable for segmentation network.
- In order to compute MI irrespective of the spatial dimensions of X_1 and X_2 transform Y into $\tilde{Y}$ through the vectorization network.

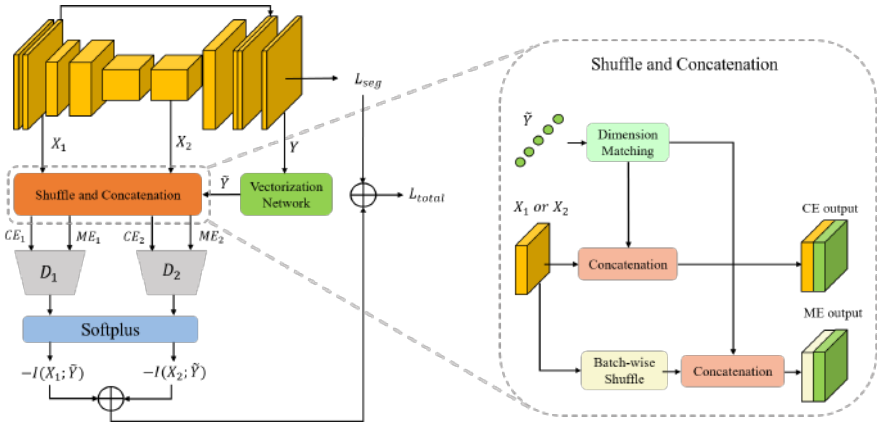

Fig. 4. The representation learning framework for semantic segmentation network.

- Based on the loss function using the computed MI, learn the segmentation network.

4 Experiments

We performed three experiments to verify the proposed system. First, we evaluated the accuracy of pupil center detection by the proposed method. The quantitative evaluation metric of accuracy is defined by Eq. (5).

$$e = \frac{max(d_l, d_r)}{d} \tag{5}$$

where d_r,d_l denote Euclidean distance between the detected pupil center and Ground Truth (GT) in the right and left eyes, respectively, and d denotes Euclidean distance between the GT centers of two eyes calculated. In addition, FLoating point Operations Per second(FLOPs), the number of parameters, and latency were measured to evaluate the light-weighting level of the face detection network and glasses removal network. Second, the qualitative test was performed. Third, an ablation study was conducted using the fine level index to evaluate each module's contribution in performance. In order to make a fair comparison, we compared the proposed method with some prior knowledge-based methods and other context-based methods, respectively.

Except landmark extractor and VGG-16, we trained other models from scratch. We used five-point landmark extractor provided by dlib[29] and VGG-16 pre-trained on ImageNet.

4.1 Datasets and Experimental Environment

We used four public datasets : WIDER FACE[30] BioID[31], GI4E[32], Talking Face Video[33]. Also, we collected a dataset to train the glasses classifier and

glasses removal network through web crawling. More detail information of each dataset is as follows.

BioID. This is a dataset consisting of 1,521 low-resolution images from various subjects. The image resolution is 384x286. Images in this dataset are negatively affected by varying illumination, head-poses and occlusion by glasses, and even include eyes closed. That is, a very challenging dataset.

GI4E. It is a dataset consisting of 1,236 images from various subjects. Most of the images in this dataset have a larger resolution (800x600) than BioID and consist of frontal face images.

Talking Face Video(TFV). It is a dataset consisting of a total of 5000 images of just one subject. The image resolution is 720 x 576. This dataset was taken while the subject was talking, including various head poses and eyes closed.

WIDER FACE. It is a dataset with annotated face location information of people in various event images. The dataset consists of 393,703 face images with various scales, head poses and occlusions.

Customized Dataset. We collected 1,700 images of glasses wearers and 1,700 images of non-wearers through web crawling and annotated them.

We composed whole training datasets and evaluation datasets as follows. We composed a training dataset for the proposed face detection network by randomly selecting 12,880 images from 60 event classes in WIDER FACE dataset. And color distortion, random cropping, scale transformation, and horizontal flipping were used as data augmentations for face detection network training. In case of glasses classifier and glasses removal network, we used 3,224 images of the customized dataset as a training dataset and 176 images as a validation dataset. For the segmentation network, we constructed the training datasets by cropping the eye region (R_i) that satisfies Eq.(6) using the label data of each image.

$$R_i = \{z \mid \| z - o^i \|_1 \leq 3l\}, i \in \{Right, Left\}, z \in \mathbb{R}^2 \tag{6}$$

where o_i is the midpoint between the two ends of eye, and l is $1/2$ of the distance between the two ends of the eye. Meanwhile, we employed two different dataset compositions for the segmentation network. Firstly, due to lack of dataset, we integrated BioID and GI4E dataset. And we equally split the integrated dataset into five-folds. In other words, we used 80% data of the integrated dataset as a training dataset and evaluated the proposed method by using the other data for each fold. Secondly, we entirely used the integrated dataset for training dataset and evaluated the segmentation network on TFV dataset. Color jittering was used as data augmentation for training of the segmentation network. Note that the TFV is only used for evaluation. The more details of implementation can be found in the supplementary material.

Meanwhile, each module in the PCT system was learned individually. A hardware environment consisting of one NVIDA GTX 1070 Ti GPU and one Intel i7-8700 CPU was used.

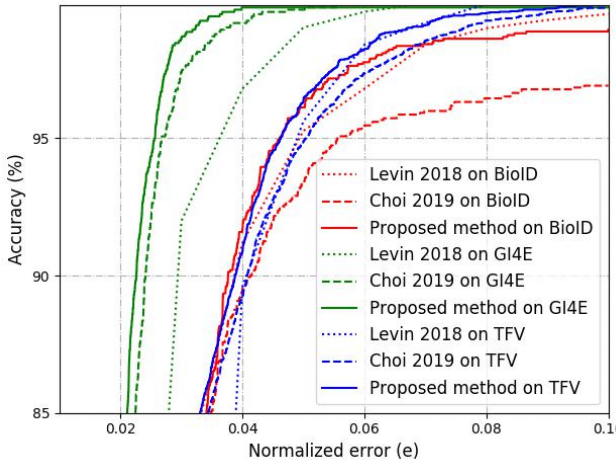

Fig. 5. Accuracy vs. the normalized error (e) for BioID, GI4E and TFV datasets.

Fig. 6. Qualitative comparison between [13] (left) and ours (right) for BioID dataset.

4.2 Quantitative results

The proposed method and the SOTA methods were compared in terms of three accuracy levels for BioID, GI4E and Talking Face Video datasets. All quantitative results in Table 1 and Table 2 were acquired by five-times experiments and the figures in Tables are the average values. The first accuracy level of $e \leq 0.05$ means that the estimated pupil center position is within pupil radius, which is also called the fine level because it reflects the highest precision of pupil center detection. The middle level of $e \leq 0.1$ indicates that the estimated pupil center position is within iris radius. Finally, the coarsest level of $e \leq 0.25$ means that the estimated pupil center position is within eye radius.

Table 1 shows the experimental results for the BioID dataset. In terms of fine level ($e \leq 0.05$), the proposed method showed outstanding accuracy of 96.71% ($\pm$0.05%). This is 1.44% higher than the best SOTA method[4]. How-

Table 1. Comparison results for BioID dataset.

Method	Category	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Tian2016[5]	Prior	93.93%	98.22%	NA
Vater2016[6]	Prior	89.48%	94.85%	NA
Zhang2016[7]	Prior	85.66%	93.68%	99.21%
Kacete2016[8]	Prior	91.30%	97.90%	99.6%
Ahuja2016[9]	Prior	92.06%	97.96%	100%
Cai2018[10]	Prior	92.80%	NA	NA
Xiao2018[11]	Prior	94.35%	98.75%	99.80%
Levin2018[4]	Prior	95.27%	99.52%	100%
Gou2017[3]	Context	91.20%	99.40%	99.80%
Choi2019[13]	Context	93.30%	96.91%	100%
Xia2019[12]	Context	94.40%	**99.90**%	100%
Ours	Context	**96.71**%	98.95%	**100**%

Table 2. Comparison results for GI4E and TFV datasets.

GI4E dataset				
Method	Category	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Zhang2016[7]	Prior	97.90%	99.60%	NA
Cai2018[10]	Prior	99.50%	NA	NA
Xiao2018[11]	Prior	97.90%	100%	100%
Levin2018[4]	Prior	99.03%	99.92%	100%
Gou2017[3]	Context	98.30%	99.80%	NA
Xia2019[12]	Context	99.10%	**100**%	100%
Choi2019[13]	Context	99.60%	99.84%	100%
Ours	Context	**99.84**%	99.84%	**100**%
TFV dataset				
Ahuja2016[9]	Prior	94.78%	99.00%	99.42%
Xiao2018[11]	Prior	91.24%	97.56%	99.96%
Levin2018[4]	Prior	95.62%	99.88%	99.98%
Choi2019[13]	Context	95.18%	99.72%	100%
Ours	Context	**96.38**%	**100**%	**100**%

Table 3. Complexity comparision in terms of FLOPs, the number of parameters and latency.

	Face Detection Network			Glasses Removal Network	
Method	Choi2019[13]	Faceboxes[26]	Ours	Choi2019[13]	Ours
FLOPs	2.39G	0.09G	0.15G	245G	59G
Parameters	2.6M	1M	1.74M	11M	3.2M
Latency	13ms	10ms	7ms	10ms	4ms

ever, we could see the proposed method gives a little bit lower performance in case of $e \leq 0.1$. This is because other SOTA methods were evaluated by using clean face image through annotated data. However, unlike [3, 4, 12], the proposed

method is evaluated with a face image predicted by the proposed face detector. In other words, the inaccuracy of the face detection is included in the result of the proposed method. For the GI4E dataset, the proposed method showed the SOTA performance of 99.84%(±0.00%) in terms of fine level (see Table 2). Even for TFV dataset, the proposed method showed the best performance for fine level by 0.76% (±0.12%). Fig. 5 shows accuracy curves of the proposed method, [4] and [13] for BioID, GI4E and TFV datasets. Fig. 5 illustrates that the proposed method is more accurate than other methods under the fine level range ($e \leq 0.05$).

On the other hand, in order to quantitatively verify the complexity of the proposed PCT system, the FLOPs, the number of parameters, and the latency of the face detection network and the glasses removal network were measured. And the measured values were compared with those of [13] and [26]. In Table 3, the number of FLOPs and parameters of the proposed face detection network increased, but its latency was lower than that of [26]. This phenomenon can be attributed to the bottleneck of the concatenation structure of the inception module. Note that the FLOPs and parameters of the proposed glasses removal network decreased to only 24.1% and 29.1% of [13]. The latency of the proposed glasses removal network also decreased significantly to around 40% of [13]. As a result, the total latency of the proposed PCT system amounts to about 19ms, which is enough for real-time operation on general purpose computers.

4.3 Qualitative Results

This section qualitatively compared the proposed method and [13] (see Fig. 6). For this experiment, the BioID dataset was used. We could observe that the proposed method provides closer result to the actual pupil center. We also qualitatively verified the validity of the PCCL proposed in Section 3.3. Fig. 7 shows the results of the glasses removal network. The customized dataset was used for this experiment. Looking at the second and third rows, we could see that the PCCL significantly mitigates the blur problem. Also, in rows 3 and 4 of Fig. 7, in spite that the proposed glasses removal network is lighter than [13], the details are better preserved than [13].

4.4 Ablation study

This section further analyzes the effects of the proposed techniques on pupil center detection performance. BioID dataset was used for this experiment, and the fine level of $e \leq 0.05$ was evaluated. Since we used the five-point landmark extractor provided by dlib[29] to ensure low latency, we increased the size of the bounding box horizontally by about 10 pixels to extract the exact landmarks. For fair comparison, the size of the bounding box obtained from the face detection network of [13] was also increased by 10 pixels.

The experiment identifies the effects of the proposed techniques by replacing each module of [13] with the proposed technique. Table 4 shows the experimental results. The fine level accuracy of [13] was 95.39%. We used this accuracy as a

Fig. 7. Qualitative comparison with [13]'s glasses remover and ours for the customized dataset. First row is input images. Second and third rows are the proposed glasses remover's outputs without/with PCCL. Lastly, fourth row is [13]'s glasses remover's outputs.

Table 4. The effect of each module on the overall performance.

Face Detection (FD)	Glasses Removal (GR)	Mutual Information (MI)	$e \leq 0.05$
			95.39%
✓			95.59%
	✓		95.79%
✓	✓		95.99%
✓	✓	✓	96.71%

baseline in the following experiment. When the face detection network of [13] is replaced with the proposed network (FD), the performance increases to 95.59% (+0.2%). In case of changing the glasses removal network to the proposed method (GR), the overall performance was 95.79% (+0.4%). If both face detection and glasses removal networks of [13] were modified with the proposed techniques (FD + GR), the detection accuracy became 95.99% (+0.6%). Finally, when all the modules including the segmentation network were replaced with the proposed methods (FD + GR + MI), that is, the proposed PCT system itself showed the detection accuracy of 96.71% (+1.32%). Meanwhile, we investigated practical effect of the proposed representation learning on the segmentation network (see Fig. 8). As shown in Fig. 8 (b) and (d), the proposed representation learning using MI did not practically affect on output shapes. However, as shown in Fig. 8 (a) and (c), we can observe that the proposed representation learning provides the segmentation network with additional location information for accurate pupil

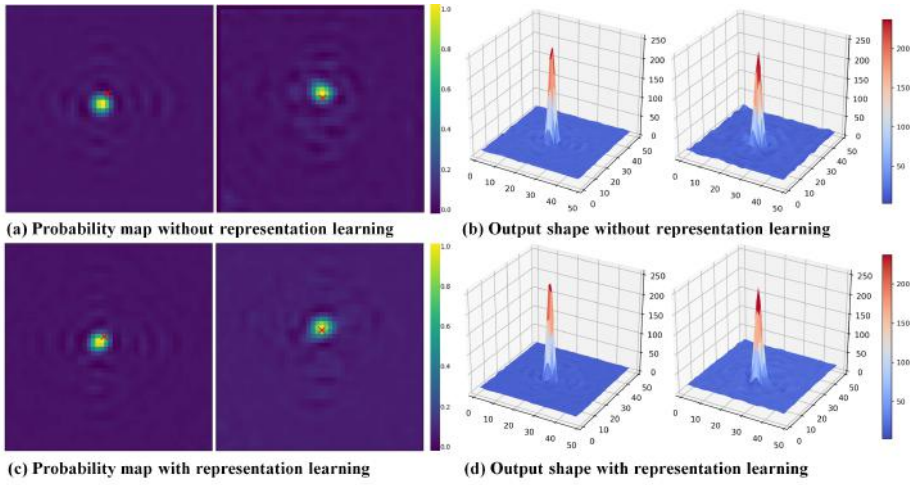

Fig. 8. (a) and (c) are the probability maps of pupil center locations. In (a) and (c), Red 'x' marks are ground truth locations.(b) and (d) are the shapes of segmentation network outputs. Here, BioID dataset was used.

center detection. To sum up with, all of the proposed modules provide significant performance improvements, and the segmentation network combined with the proposed representation learning has the greatest effect on pupil center detection performance without additional cost during inference.

5 Conclusions

This paper proposes pupil center detection methods for high accuracy and light weight methods to secure real-time operation. The proposed representation learning can provide an additional information into the segmentation network for accurate pupil center detection by using mutual information. Also, we designed a low latency face detection network using a non-local block and a lightweight glasses removal network that provides good image quality by using self-attention block and perceptual loss. Experimental results show that the proposed scheme has a low latency of 19ms per frame with state-of-the-art accuracy in fine level index. The proposed real-time PCT system is expected to be effectively used in systems such as AR, VR, and hologram.

Acknowledgements. This work was supported by 'The Cross-Ministry Giga KOREA Project' grant funded by the Korea government (MSIT) [1711093798, Development of full-3D mobile display terminal and its contents] and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [2020-0-01389, Artificial Intelligence Convergence Research Center (Inha University)].

References

1. Li Jianfeng and Li Shigang. Eye-model-based gaze estimation by rgb-d camera. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2014.
2. Fabian Timm and Erhardt Barth. Accurate eye centre localisation by means of gradients. 2011.
3. Chao Gou, Yue Wu, Kang Wang, Kunfeng Wang, Fei-Yue Wang, and Qiang Ji. A joint cascaded framework for simultaneous eye detection and eye state estimation. *Pattern Recognition*, 67:23–31, 2017.
4. Alex Levinshtein, Edmund Phung, and Parham Aarabi. Hybrid eye center localization using cascaded regression and hand-crafted model fitting. *Image and Vision Computing*, 71:17–24, 2018.
5. Dong Tian, Guanghui He, Jiaxiang Wu, Hongtao Chen, and Yong Jiang. An accurate eye pupil localization approach based on adaptive gradient boosting decision tree. In *2016 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2016.
6. Sebastian Vater and Fernando Puente León. Combining isophote and cascade classifier information for precise pupil localization. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 589–593. IEEE, 2016.
7. Wenhao Zhang, Melvyn L Smith, Lyndon N Smith, and Abdul Farooq. Eye center localization and gaze gesture recognition for human–computer interaction. *JOSA A*, 33(3):314–325, 2016.
8. Amine Kacete, Jerome Royan, Renaud Seguier, Michel Collobert, and Catherine Soladie. Real-time eye pupil localization using hough regression forest. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016.
9. Karan Ahuja, Ruchika Banerjee, Seema Nagar, Kuntal Dey, and Ferdous Barbhuiya. Eye center localization and detection using radial mapping. In *2016 IEEE International Conference on image processing (ICIP)*, pages 3121–3125. IEEE, 2016.
10. Haibin Cai, Bangli Liu, Zhaojie Ju, Serge Thill, Tony Belpaeme, Bram Vanderborght, and Honghai Liu. Accurate eye center localization via hierarchical adaptive convolution. In *BMVC*, page 284, 2018.
11. Feng Xiao, Kejie Huang, Yue Qiu, and Haibin Shen. Accurate iris center localization method using facial landmark, snakuscule, circle fitting and binary connected component. *Multimedia Tools and Applications*, 77(19):25333–25353, 2018.
12. Yifan Xia, Hui Yu, and Fei-Yue Wang. Accurate and robust eye center localization via fully convolutional networks. *IEEE/CAA Journal of Automatica Sinica*, 6(5):1127–1138, 2019.
13. Jun Ho Choi, Kang Il Lee, Young Chan Kim, and Byung Cheol Song. Accurate eye pupil localization using heterogeneous cnn models. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2179–2183. IEEE, 2019.
14. P Viola and MJ Jones. «robust real-time face detection», international journal of computer vision, vol. 57, no. 2. 2004.
15. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
16. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

17. Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.
18. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
19. Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
20. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
21. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
22. R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
23. Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.
24. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
25. Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.
26. Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. Faceboxes: A cpu real-time face detector with high accuracy. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2017.
27. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
28. Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.
29. Davis E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, December 2009.
30. Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.
31. Oliver Jesorsky, Klaus J Kirchberg, and Robert W Frischholz. Robust face detection using the hausdorff distance. In *International conference on audio-and video-based biometric person authentication*, pages 90–95. Springer, 2001.
32. Arantxa Villanueva, Victoria Ponz, Laura Sesma-Sanchez, Mikel Ariz, Sonia Porta, and Rafael Cabeza. Hybrid method based on topography for robust detection of iris center and eye corners. *ACM Trans. Multimedia Comput. Commun. Appl.*, 9(4), August 2013.
33. Dijana Petrovska-Delacrétaz, Sylvie Lelandais, Joseph Colineau, Liming Chen, Bernadette Dorizzi, M Ardabilian, E Krichen, M-A Mellakh, A Chaari, S Guerfi,

et al. The iv 2 multimodal biometric database (including iris, 2d, 3d, stereoscopic, and talking face data), and the iv 2-2007 evaluation campaign. In *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, pages 1–7. IEEE, 2008.