algorithms

MDPI

Article

The Effect of Different Deep Network Architectures upon CNN-Based Gaze Tracking

Hui-Hui Chen [1], Bor-Jiunn Hwang [1,*], Jung-Shyr Wu [2] and Po-Ting Liu [1]

[1] Department of Computer and Communication Engineering, Ming Chuan University, Taoyuan 333, Taiwan; huichen@mail.mcu.edu.tw (H.-H.C.); 07166135@me.mcu.edu.tw (P.-T.L.)

[2] Department of Communication Engineering, National Central University, Taoyuan 320, Taiwan; jswu@ee.ncu.edu.tw

* Correspondence: bjhwang@mail.mcu.edu.tw

Received: 12 March 2020; Accepted: 17 May 2020; Published: 19 May 2020

Abstract: In this paper, we explore the effect of using different convolutional layers, batch normalization and the global average pooling layer upon a convolutional neural network (CNN) based gaze tracking system. A novel method is proposed to label the participant's face images as gaze points retrieved from eye tracker while watching videos for building a training dataset that is closer to human visual behavior. The participants can swing their head freely; therefore, the most real and natural images can be obtained without too many restrictions. The labeled data are classified according to the coordinate of gaze and area of interest on the screen. Therefore, varied network architectures are applied to estimate and compare the effects including the number of convolutional layers, batch normalization (BN) and the global average pooling (GAP) layer instead of the fully connected layer. Three schemes, including the single eye image, double eyes image and facial image, with data augmentation are used to feed into neural network to train and evaluate the efficiency. The input image of the eye or face for an eye tracking system is mostly a small-sized image with relatively few features. The results show that BN and GAP are helpful in overcoming the problem to train models and in reducing the amount of network parameters. It is shown that the accuracy is significantly improved when using GAP and BN at the mean time. Overall, the face scheme has a highest accuracy of 0.883 when BN and GAP are used at the mean time. Additionally, comparing to the fully connected layer set to 512 cases, the number of parameters is reduced by less than 50% and the accuracy is improved by about 2%. A detection accuracy comparison of our model with the existing George and Routray methods shows that our proposed method achieves better prediction accuracy of more than 6%.

Keywords: gaze tracking; convolution neural network; batch normalization; global average pooling layer

1. Introduction

Gaze tracking can help understand cognitive processes and emotional state, and has been applied in many fields, such as medicine, Human-Computer Interaction (HCI), and e-learning [1–3]. The techniques of gaze tracking are classified into two methods, model-based and appearance-based [4]. First, the model-based method mainly uses the near-infrared light device to track the pupil position and the designed algorithm to estimate the gaze points which usually require expensive hardware [5]. A simple video-based eye tracking system was developed with one camera and one infrared light source to determine a person's point of regard (PoR) [6], assuming the location of features in the eye video is known. Zhu et al. [7] used the dynamic head compensation model to solve the effect of head movement for estimating the gaze movement. Zhou et al. [8] used the Kinect sensor to detect the three-dimensional coordinates of the head movement. Then the gaze is calculated by

using the estimated eye model parameters acquired from gradients iris center localization method, geometric constraints-based and Kappa angle calculation method.

Second, the appearance-based method mainly uses the technology of machine learning to gain the features of a large number of input samples, such as eye, eyes or face, and then used the learning model to predict the gaze [9]. Wu et al. [10] proposed two procedures to estimate the gazing direction. First, the eye region is located by modifying the characteristics of the Active Appearance Model. Then the five gazing direction classes are predicted by employing the Support Vector Machine (SVM). Because deep learning (DL) and Convolution Neural Network (CNN) has a prominent performance in computer vision, there are some studies that are used to improve the accuracy of eye movement prediction considering as a regression task [11–15] or a classification task [11,16]. The authors applied the convolutional neural network and trained a regression model in the output layer for gaze estimation [11]. Krafka et al. developed the GazeCapture for gaze prediction, the first large-scale eye tracking dataset captured via crowdsourcing and iTracker, a convolutional neural network, is trained [12]. Wang et al. [13] proposed estimating the gaze angle by dividing the screen area into 41 points and random forest regression is used. In [14], a convolutional neural network model was introduced as a regression problem with finding a gaze angle. This model has low computational requirements, and effective appearance-based gaze estimation was performed on it. Zhang et al. [15] proposed a gaze estimation method in which a CNN utilizes the full face image as input with spatial weights on the feature maps. Based on a CNN classification task, the eye image is input to the multi-scale convolutional layer for depth feature extraction. The gaze direction classification is treated as a multi-class classification problem [16]. The left and right eyes are trained separately by two convolutional neural networks to classify the gaze in seven directions. The scores from both the networks are used to obtain the class labels. Zhang et al. [11] created the MPIIGaze dataset that contains 213,659 images collected from 15 participants. First, the head rotation angle and eye coordinates of the facial image are obtained through the head pose model and facial feature model. Then, the multimodal CNN is used to learn the mapping from the head poses and eye images to gaze directions in the camera coordinate system. Zhang et al. [17] did not directly estimate the gaze angle, but introduced a method to divide human gaze into nine directions, and established a convolutional neural network model to estimate directions for a screen typing application.

Many studies predicted the coordinates of the gaze point by the regression method. The accuracy of gaze estimation is expressed as the error of gaze direction or coordinate angle (degree). Although both methods are useful, we observed the incorrect predictions and found that the incorrect estimated class is usually near the correct class. Therefore, we infer that when the user views the edges in two adjacent blocks, it may cause the CNN model to make an incorrect estimation. In general, users watch videos or animations and are interested in objects. For establishing a relationship between an area-of-interest and an interesting object, according to our previous work [18], the ratio of each area-of-interest is the probability of the interest in an object. Thus, when the amount of probability is calculated, one can estimate the amount of attention paid to the object of interest. Even if it is misidentified as a neighboring block, it may still correspond to the same object. Especially when the block area is relatively small, the misjudgment will be higher, and the influence will be reduced by using the classification method. Thus, this paper treats the eye gaze estimation task as a classification problem.

By adding convolution and pooling layers compared to traditional neural networks [19], the network can maintain the shape of the image information and reduce parameters. The convolution neural network for the research of gaze tracking can bring great results since convolutional neural networks capture subtler features through deep networks and make a robust model. Several papers have clearly helped improve CNN performance [20–23]. Ioffe et al. [20] proposed the method of batch normalization (BN) to overcome the problem of hard to train models with saturating nonlinearities by making normalization a part of the model architecture and performing the normalization for each training mini-batch. The BN can improve the learning speed, reduce the dependence on initial parameters, and eliminate the need for Dropout layer in some cases. GoogleNet [21] is proposed by

Szegedy et al. who increased the depth and width of the network while keeping the computational budget constant, to perform multi-scale processing on images and greatly reduced the amount of model parameters. Simonyan et al. [22] studied very deep convolutional networks by fixing other parameters and adding more convolutional layers. They used very small (3×3) convolution filters in all layers. This is beneficial for the classification accuracy, and good performance on the ImageNet challenge dataset can be achieved. Lin et al. [23] proposed using the global average pooling (GAP) layer, over feature maps in the classification layer, instead of the fully connected layer. The results have shown that it is easier to interpret and less prone to overfitting.

The objective of most of the above research is concentrated on recognizing large sized images. In the case of gaze tracking, the input image usually is the eye or face, however, few studies focus on small sized image with relatively few features. The images (static stimuli) are used to construct the training dataset [11,12]. However, most visual behavior of many people involves watching dynamic stimuli such as movies from YouTube and Netflix. In this paper, we propose a gaze tracking method with feeding different types of images, including the single eye, double eyes and face based on CNN. Additionally, gaze prediction is considered a classification problem by dividing the screen into blocks to label the gaze. In order to obtain the fed images with gaze label for feature learning in a way that is close to the viewer's visual behavior, the data collection involves participants watching the videos. Performing this way is closer to the actual visual behavior of the viewer rather than letting the participants watch specific screen blocks such as [13]. The training images then are fed into the convolution neural network for training prediction model, performing several experiments by adjusting network parameters or architecture to explore the performance of the model. The highlights of the paper are shown below:

- We propose a framework for eye gaze classification to explore the effect by using different convolutional layers, batch normalization and the global average pooling layer.
- We propose three schemes, including the single eye image, double eyes image and facial image to evaluate the efficiency and computing complexity.
- We propose a novel method to build a training dataset, namely watching videos, because this is closer to the viewer's visual behavior.

The remainder of this paper is organized as follows. Section 2 first gives the details of the proposed method, different convolutional layers, batch normalization and the global average pooling layer. The numerical analysis and performance comparison are given in Section 3. Finally, we will draw conclusions in Section 4.

2. Materials and Methods

In order to build a complete gaze tracking system, the proposed method comprises of four steps, collecting data, preprocessing data, training model and evaluating and testing, as shown in Figure 1. The first step is collecting data. The system detects the participant's facial image while watching the videos and an eye tracker are adopted to label the gaze point. The second step is preprocessing data. The collected images are firstly checked and cropped, and data augmentation is performed to increase the amount of training samples. The third step is the training model. The system uses the fed images, the single eye, the double eyes and the face, into proposed network for training. Finally, through evaluating and testing by adjusting network parameters and architecture, the optimized model can be acquired.

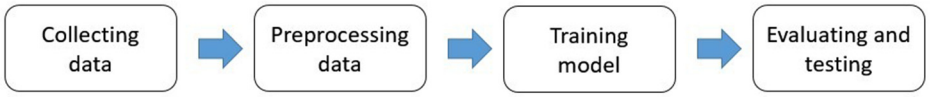

Figure 1. Flowchart of process.

2.1. Collecting Data

The 22 inch with 16:9 screen is used as the video playback device, Logitech C920r HD Pro Webcam and Tobii Eye Tracking 4C as the experimental tools for collecting data. The accuracy of the eye tracker for gaze tracking provides good evidence for experimentally collected data. Under normal circumstances, the distance between the screen and the participant is about 70 cm. The webcam is placed above the screen and the eye tracker is placed under the screen, as shown in Figure 2. The participant can swing his head freely within the range detectable by the webcam. In this way, the most real and natural viewing information can be collected without too many restrictions. Each participant watches two videos selected randomly from six animated videos with length between 245–333 s, as shown in Table 1.

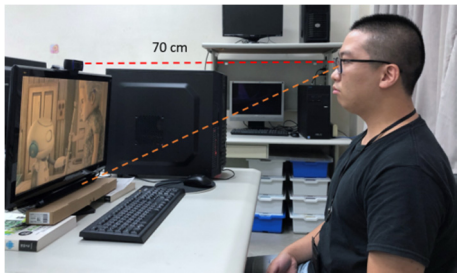

Figure 2. Schematic diagram of the experimental environment.

Table 1. The details of six animated videos.

Title	Resolution (Pixel)	Length (sec)
Reach [1]	1920 × 1080	193
Mr Indifferent [2]	1920 × 1080	132
Jinxy Jenkins & Lucky Lou [3]	1920 × 1080	193
Changing Batteries [4]	1920 × 1080	333
Pollo [5]	1920 × 1080	286
Pip [6]	1920 × 1080	245

[1] https://www.youtube.com/watch?v=OL5PVmeQApM. [2] https://www.youtube.com/watch?v=qLGNj-xrgvY. [3] https://www.youtube.com/watch?v=OuJ4BBQ0nhc. [4] https://www.youtube.com/watch?v=O_yVo3YOfqQ. [5] https://www.youtube.com/watch?v=ExP3VGSPzrU&feature=youtu.be&fbclid. [6] https://www.youtube.com/watch?v=07d2dXHYb94.

The flowchart of collecting data is shown in Figure 3. The eye tracker executes a calibration procedure prior to collecting data. The OpenCV [24] and eye tracker are used to detect the participant's face and gaze, respectively. Finally, facial images are labeled as gaze.

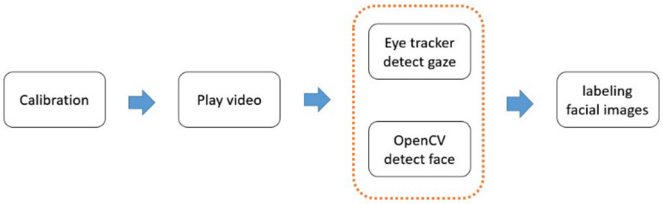

Figure 3. The flowchart of collecting data.

To reduce the complexity of predicting the gaze as a classification method, and to achieve a balance in the size of the block area, this paper refers to the design composition method and the range of the focus area [11] to divide the screen into blocks, as shown in Figure 4. The screen contains 37 blocks in total. The block numbers 0 to 24 are located in the central area of the screen, and each block size is of 4.8 cm × 2.7 cm. The block numbers 25 to 36 are located in the peripheral area of the screen, and each

block size is 12 cm × 6.75 cm. Therefore, the labeled gazes in related to the facial image are categorized into 37 blocks to train the model.

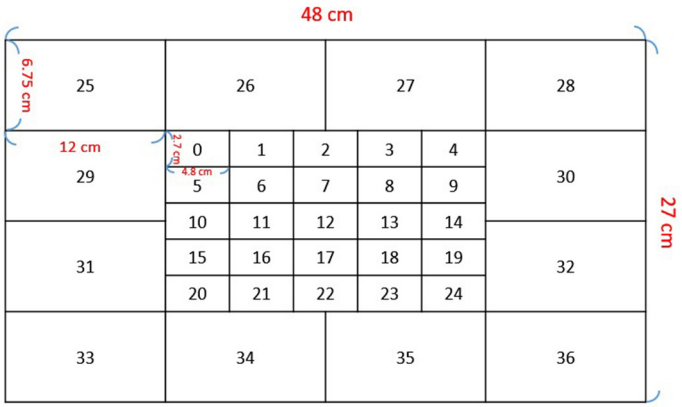

Figure 4. Screen block diagram.

2.2. Preprocessing Data

The collected images are checked firstly to eliminate unsuitable images. Then, the cropping procedure is performed to crop out the required image size used as a training sample, 32 × 32 pixels for single eye scheme, 96 × 32 pixels for double eyes scheme and 200 × 200 pixels for face scheme.

Collecting a lot of data to acquire a good training model is a difficult task; therefore, data augmentation is applied. The data augmentation has the benefit for reducing the probability of overfitting. Two methods of data augmentation are used in this paper to evaluate the efficiency, namely brightness and noise.

2.3. Training Model

Three types of fed images and varied network architectures are applied to estimate and compare the efficiency including the number of convolutional layers (NoCL), batch normalization (BN) and the global average pooling layer (GAP) instead of the fully connected layer, as shown in Figure 5. Additionally, the amount of the fully connected layers (FC) is adjusted to evaluate the efficiency. In the experimental description, for example, FC1024 represents the neuron parameter amount of the fully connected layer set to 1024. As shown in Figure 6, C1 to C16 represent 16 combinations of NoCL. Taking C16 as an example, this model architecture includes 16 convolutional layers and 5 pooling layers. The BN is added after each convolutional layer and fully connected layer. Finally, the model sets a GAP instead of a fully connected layer.

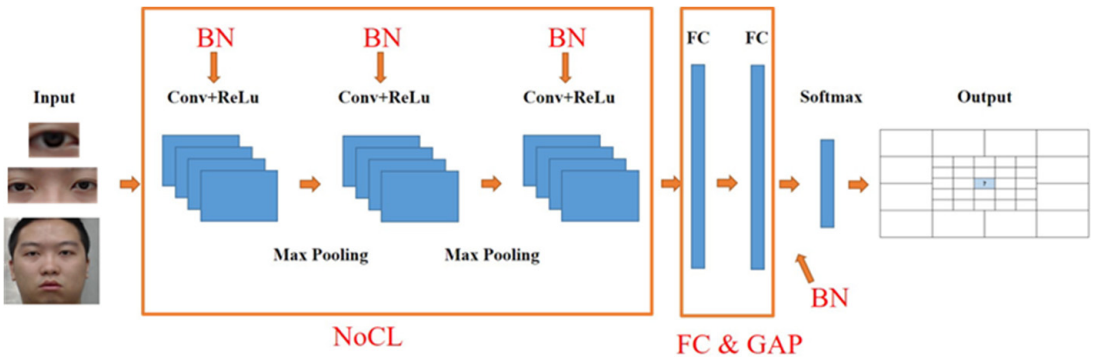

Figure 5. Parameters adjustment diagram.

		Number of Convolutional Layers															
		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
Input		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
depth=64	Conv1_1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Conv1_2		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Maxpool		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
depth=128	Conv2_1			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Conv2_2				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Maxpool				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
depth=256	Conv3_1					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Conv3_2						✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Conv3_3							✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Conv3_4								✓	✓	✓	✓	✓	✓	✓	✓	✓
Maxpool						✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
depth=512	Conv4_1									✓	✓	✓	✓	✓	✓	✓	✓
	Conv4_2										✓	✓	✓	✓	✓	✓	✓
	Conv4_3											✓	✓	✓	✓	✓	✓
	Conv4_4												✓	✓	✓	✓	✓
Maxpool										✓	✓	✓	✓	✓	✓	✓	✓
depth=512	Conv5_1													✓	✓	✓	✓
	Conv5_2														✓	✓	✓
	Conv5_3															✓	✓
	Conv5_4																✓
Maxpool														✓	✓	✓	✓
FC		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
FC		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
FC		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Softmax		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Figure 6. Comparison of the number of convolutional layers.

2.4. Evaluating and Testing

There are 5 experiments designed to evaluate the performance, as shown in Table 2. The experiment 1 will evaluate the adjustment of NoCL. The experiment 2 will use the parameter with the highest accuracy of experiment 1 as the setting to evaluate the adjustment of the parameter FC. The experiment 3 will add BN to evaluate the adjustment of NoCL. The experiment 4 will replace the fully connected layer with a global average pooling layer to evaluate the adjustment of NoCL. The experiment 5 will use BN and GAP to evaluate the adjustment of NoCL.

Table 2. Parameter adjustment.

Experiment Number	Adjusted Parameters
1	NoCL
2	FC
3	NoCL + BN
4	NoCL + GAP
5	NoCL + GAP + BN

3. Results

The experimental results will explain the data augmentation and evaluating model. Data are collected from 22 participants resulted in a total of 83,366 facial images with labeled gazes. The data distribution of each block is shown in Figure 7. We use 3.0 GHz CPU (Intel Xeon E5-2620V4) and 1 GPU (Nvidia GTX 1080 Ti) to speed up training.

1302	2718			2621		1333
2919	1448	2545	3906	2212	1567	2411
	2246	3675	4247	3206	1650	
	1873	3622	5037	3080	1653	
2375	919	1880	3688	2292	1172	3089
	797	1314	2028	1151	776	
1582	2232			1638		1162

Figure 7. Distribution of data.

3.1. Results of Data Augmentation

In order to explore how much the model can improve the prediction results, the single eye and the face schemes are introduced as described in the evaluation. Two data augmentation methods, brightness adjustment and noise disturbance, are applied to evaluate the improvement of accuracy. Figures 8 and 9 show the results respectively of the single eye scheme and face scheme. The augmented brightness is set to adjust by −20%, −10%, 0%, 10%, 20%. The noise disturbance is changed randomly from 0 to 5% of the image pixels to white points. The experimental parameters are set as NoCL to C13, Lr to 1×10^{-5}, FC512, and without BN. The accuracy of the brightness adjustment is improved by 35.7% and 22% respectively for the single eye scheme and for the face scheme. The accuracy of the noise disturbance is improved by 26.6% and 17% respectively for the single eye scheme and the face scheme. The accuracy of combining the brightness adjustment and noise disturbance is improved by only 31.8% and 15.6% respectively for the single eye scheme and the face scheme. The brightness dataset has the highest accuracy; therefore, the brightness adjustment will be used as the data augmentation in the following performance evaluations.

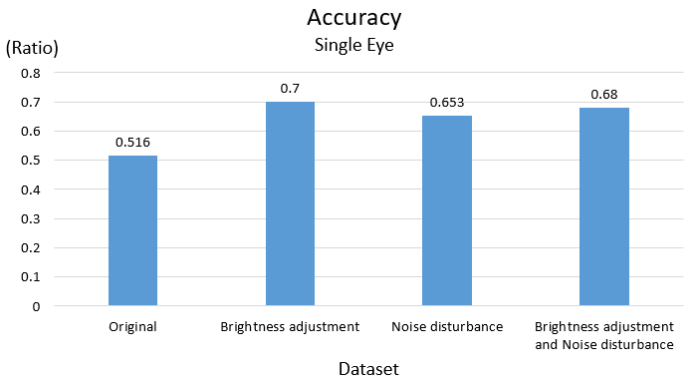

Figure 8. The accuracy of data augmentation for the single eye scheme.

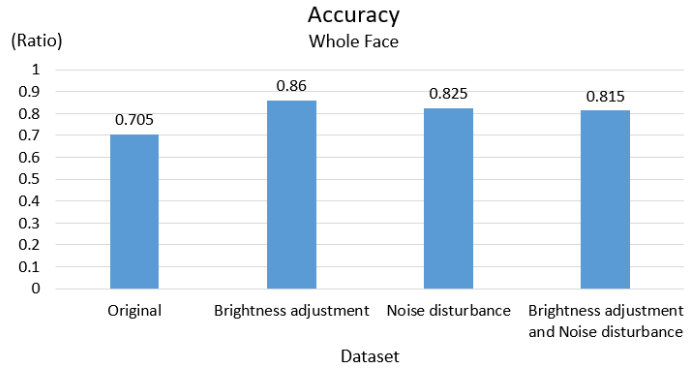

Figure 9. The accuracy of data augmentation for the face scheme.

3.2. Evaluating Model

Part of the experiments will be implemented for different schemes to explore the effect by adjusting network parameters and architecture including NoCL, Lr, FC, BN and GAP referring to Table 2.

3.2.1. Single Eye Scheme

According to Table 2, all of the experiments are selected for the single eye scheme, called Experiment 1-1, Experiment 1-2, Experiment 1-3, Experiment 1-4, Experiment 1-5, respectively.

The Experiment 1-1 is the adjusted NoCL, and other parameters are set as Lr to 1×10^{-5}, FC4096, and without BN. The results are shown in Figure 10. The accuracies of C3 to C7 are quite close, but C5 and C6 perform better (0.696). Above C8 shows a decreasing trend.

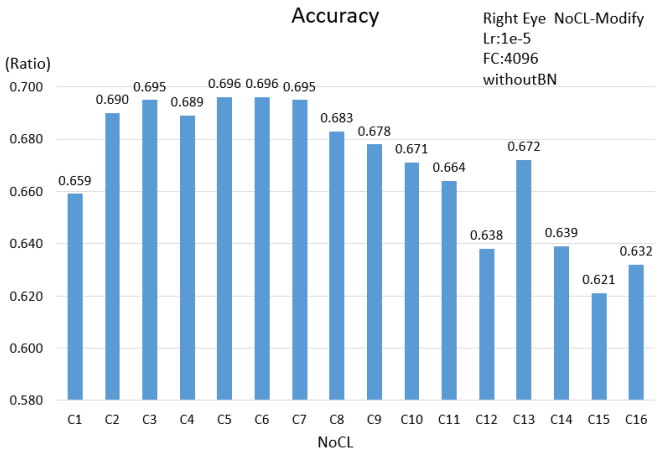

Figure 10. The accuracy of the single eye scheme for Experiment 1-1.

Because the parameters of the fully connected layer account for about 80% of the network, if the parameters can be reduced, the computing performance of the model will be improved. The Experiment 1-2 will adjust the neuron parameters of the fully connected layer and evaluate them with C6. The results are shown in Figure 11.

Although the accuracy rates of FC512, FC1024 and FC2048 are similar, the number of parameters shows multiple differences as shown in Figure 12. The number of parameters of FC512 is only 1/10 times of FC4096, and the accuracy is slightly higher than FC4096 remaining at 0.7. Therefore, the FC512 with fewer parameters is a good choice.

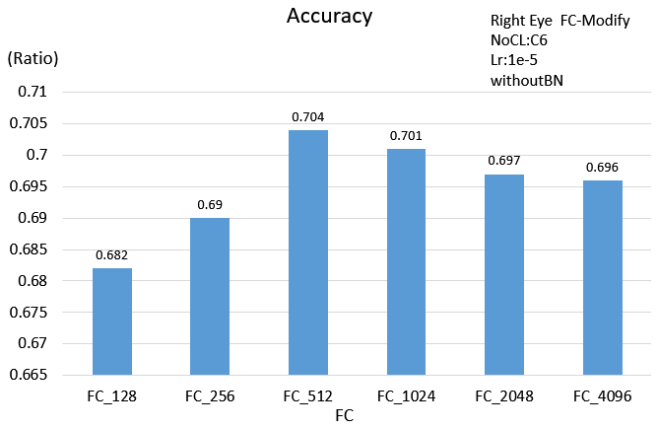

Figure 11. The accuracy of the single eye scheme for Experiment 1-2.

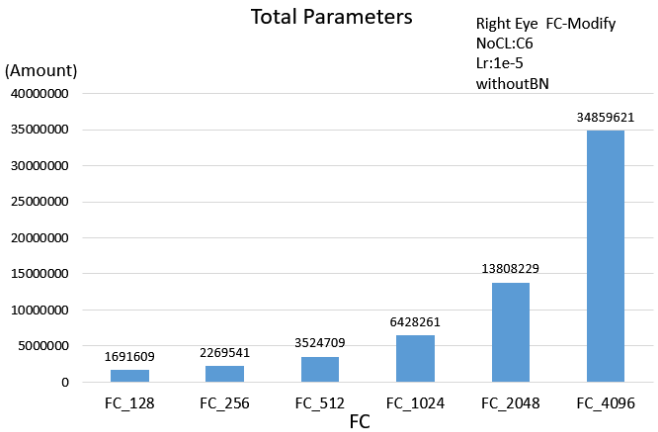

Figure 12. The number of parameters of the single eye scheme for different FC.

The benefit of adding BN is that it can make model training much more stable [20]. In this paper, we explore the effect of BN on small size of training image and a small number of network layers. Experiment 1-3 will add BN to different NoCL as model evaluation. It will take a lot of time to do all cases of the NoCL, so the evaluation will be performed at C5 to C9, which have closer accuracy. According to Figure 13, the results show that it is different from the general expectation showing a slight decrease in accuracy in cases of C5, C6 and C7. The reasons for this are that the network is not deep and the input image features are few.

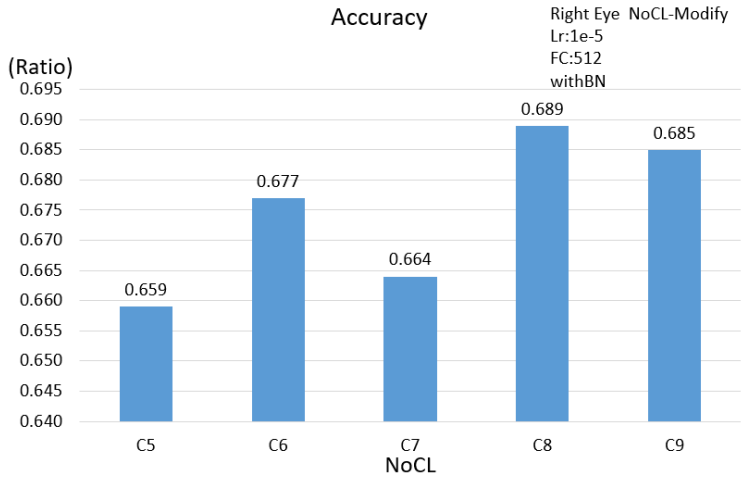

Figure 13. The accuracy of the single eye scheme for Experiment 1-3.

By replacing the fully connected layer with the global average pooling layer, a large number of parameters can be reduced, and the feature mapping can be more direct. Experiment 1-4 will evaluate the effect on the model when GAP replaces the fully connected layer. As shown in Figure 14, after replacing the fully connected layer, the amount of parameter is reduced by about 50%. Although the parameters are greatly reduced, the accuracy has a slight decrease in case of C8 compared with Figure 11, as shown in Figure 15.

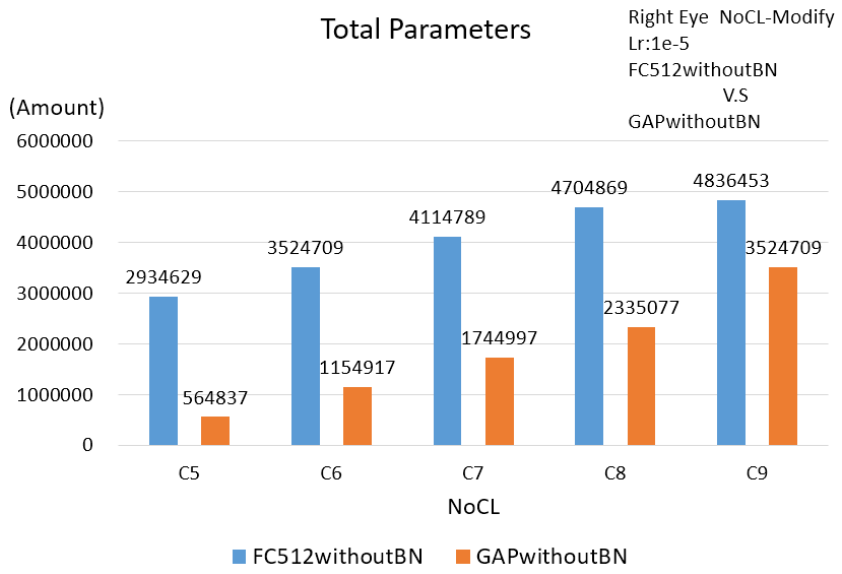

Figure 14. The number of parameters of the single eye scheme for replacing FC.

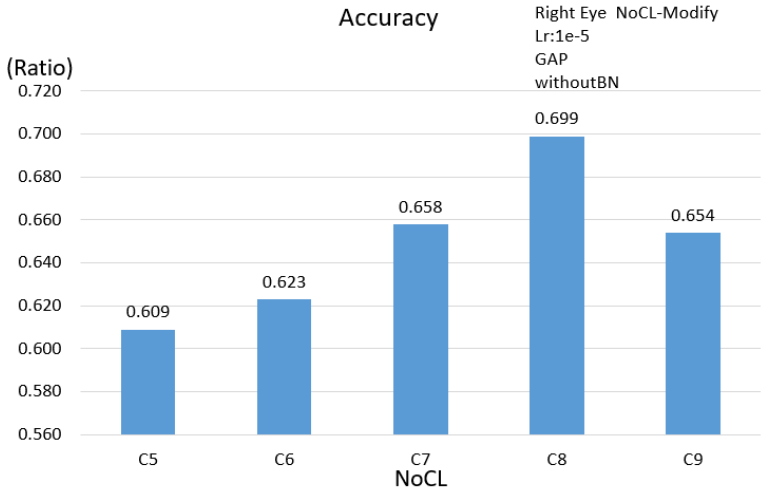

Figure 15. The accuracy of the single eye scheme for Experiment 1-4.

In Experiment 1-5, the GAP will be used to replace the FC and BN will be used to evaluate the performance of the model. The results are shown in Figure 16; compared with Figure 15, the accuracies are increased.

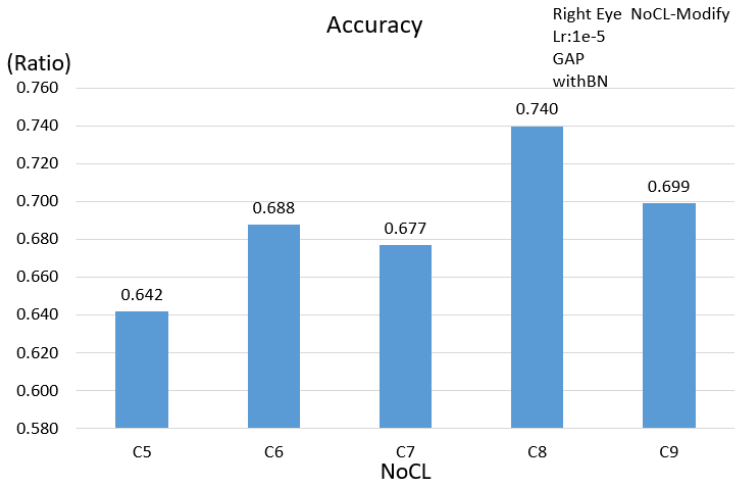

Figure 16. The accuracy of the single eye scheme for Experiment 1-5.

The accuracy of the single eye is about 0.7; however, after replacing the fully connected layer by GAP, the amount of parameter is reduced by about 50%, and the convergence speed becomes faster through adding BN. In the case of the best accuracy C8, BN and GAP are used; the execution time for a gaze estimating is 4.466 ms.

3.2.2. Double Eyes Scheme

According to Table 2, the experiments 1, 3, 4 and 5 are selected for the double eyes scheme, called Experiment 2-1, Experiment 2-2, Experiment 2-3 and Experiment 2-4, respectively.

First, under the conditions of Lr = 1×10^{-5}, FC = 512, without BN and GAP, the Experiment 2-1 is adjusted NoCL, and results show that the accuracies of C6 to C12 are all close to 0.8, as shown in Figure 17. Compared to the single eye scheme, the accuracies of the single eye scheme and the double eyes scheme are 0.704 (as shown in Figure 11) and 0.785, respectively, an increase of 11.5%. The reason for this is that the double eyes image has more features than the single eye image, and both eyes can tolerate a small amount of head swing.

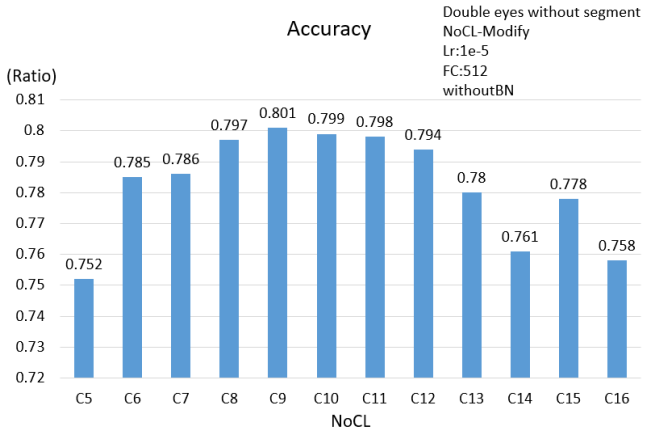

Figure 17. The accuracy of the double eyes scheme for Experiment 2-1.

The Experiment 2-2 will adjust NoCL and set Lr to 1×10^{-4} to observe the effect of adding BN for normalization, evaluated from C5 to C12. The results are shown in Figure 18. Comparing Figure 18 with Figure 17, the accuracies increase by about 3% in all cases. This result is different from the single eye scheme, shown in Figure 13, because the double eyes image has more features than the single eye image.

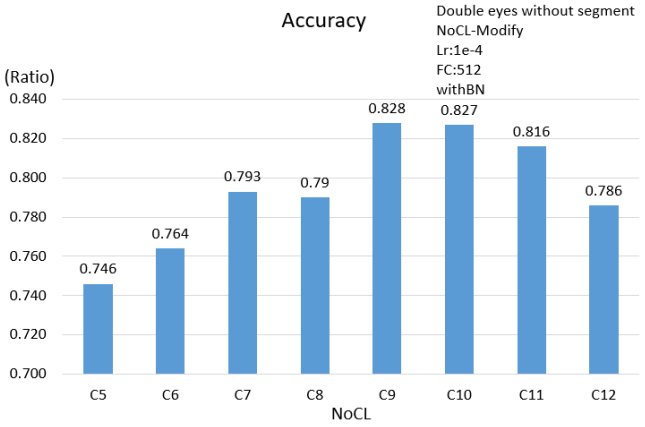

Figure 18. The accuracy of the double eyes scheme for Experiment 2-2.

Experiment 2-3 will evaluate the effect on the model when GAP replaces the fully connected layer. Comparing Figure 19 with Figures 17 and 18, the results show that the accuracies decrease. In particular, where there are more layers, the decline is greater.

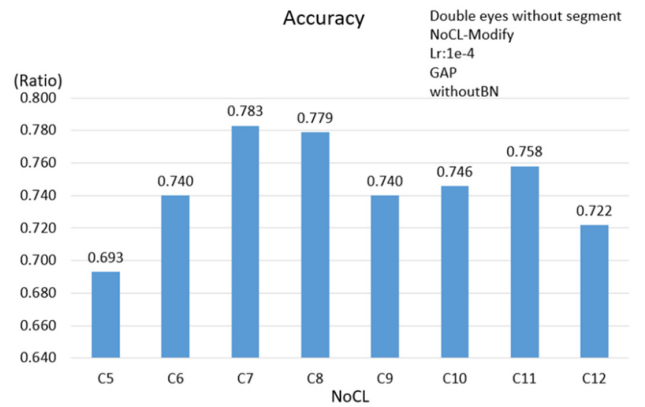

Figure 19. The accuracy of the double eyes scheme for Experiment 2-3.

In the Experiment 2-4, the GAP will be used to replace the FC and BN will be used to evaluate the performance of the model. The highest accuracy is 0.839 in the case of C8, and the execution time for a gaze estimating is 5.071 ms.

3.2.3. Face Scheme

According to Table 2, all the experiments are selected for the face scheme, called Experiment 3-1, Experiment 3-2, Experiment 3-3, Experiment 3-4 and Experiment 3-5, respectively.

Experiment 3-1 is designed to evaluate the performance by adjusting NoCL, and other parameters are set as Lr to 1×10^{-5}, FC512, and without BN. The results are shown in Figure 20. The result has the highest accuracy of 0.862 for C13. However, C5 to C8 performed the worst; the reason for this is that the network is not deep enough to learn features. In order to reduce the evaluation time, the following experiments will only be evaluated with C9 to C16.

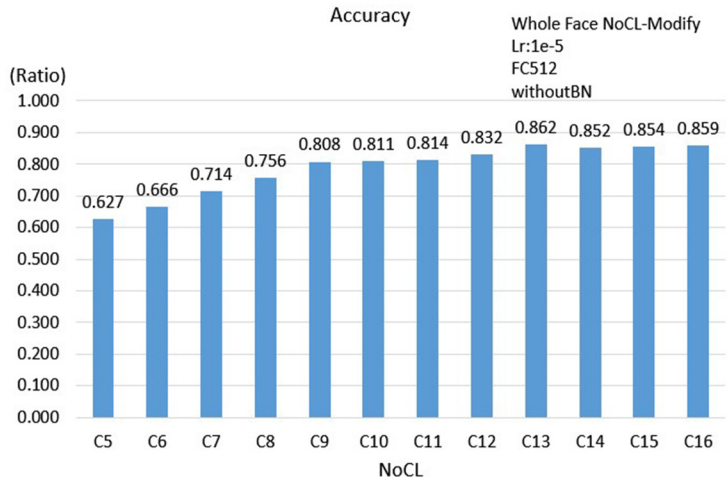

Figure 20. The accuracy of the face scheme for Experiment 3-1.

Experiment 3-2 is designed to observe the effect on the different FC based on the different types of fed images, and other parameters are set as NoCL to C13, Lr to 1×10^{-5}, and without BN. The results are shown in Figure 21. The accuracy of FC512 is only 0.001 less than FC4096. However, the parameter amount of FC512 is only 1/4 times that of FC4096, as shown in Figure 22. Therefore, in the case of the close accuracy, the FC512 with fewer parameters is a good choice.

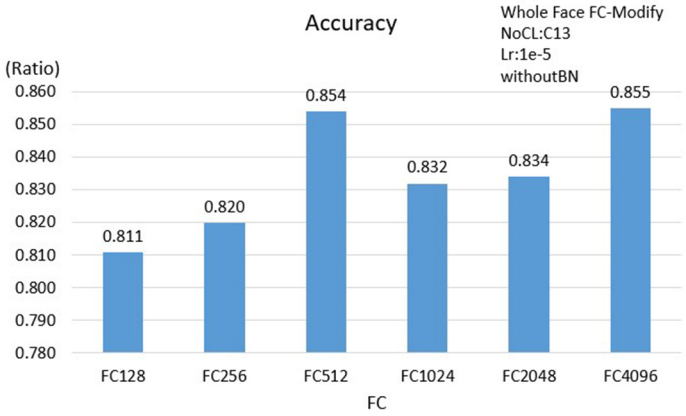

Figure 21. The accuracy of the face scheme for Experiment 3-2.

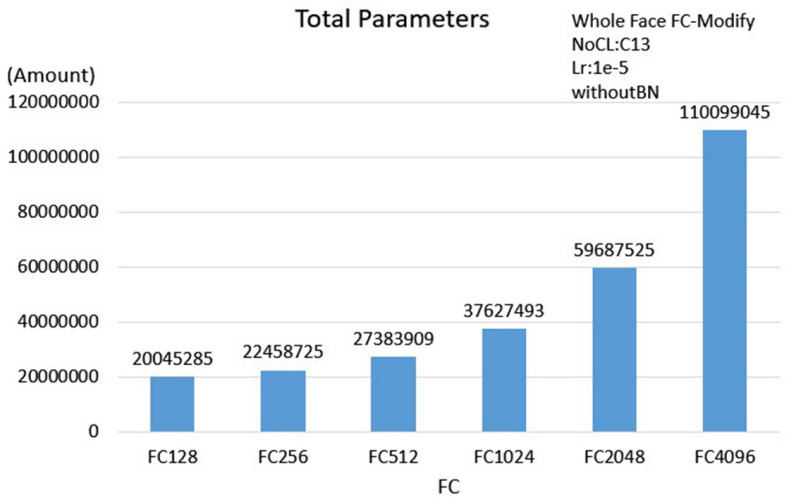

Figure 22. The number of parameters of the face scheme for Experiment 3-2.

Experiment 3-3 evaluates the accuracies for different NoCL, and other parameters are set as NoCL to C13, Lr to 1×10^{-5}, and with BN. The results are shown in Figure 23. The highest accuracy is in the case of C13, comparing with Figure 21, the accuracy is slightly improved. Additionally, comparing with the single eye scheme and the double eyes scheme, the accuracy is slightly improved due to the network being deeper.

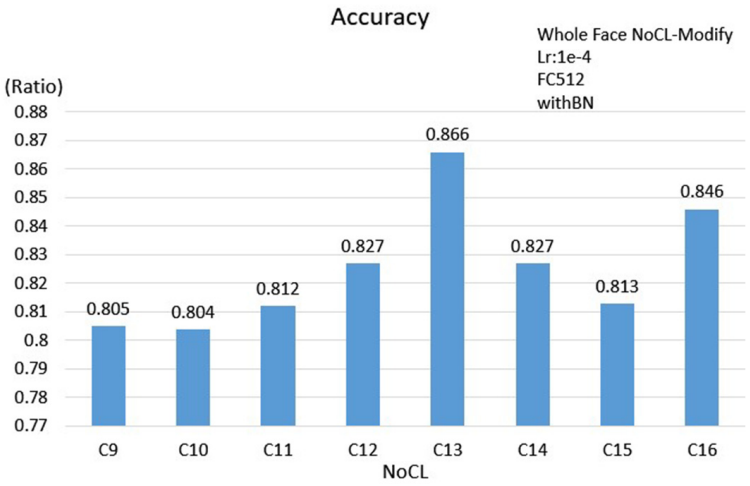

Figure 23. The accuracy of the face scheme for Experiment 3-3.

Experiment 3-4 will evaluate the effect on the model when GAP replaces the fully connected layer for different NoCL, and other parameters are set as Lr to 1×10^{-5}, without BN. The results are shown in Figure 24. The highest accuracy is C12. The number of parameters of different NoCL for comparing FC512 with GAP is described in Figure 25. For example, in the case of C12, the number of parameters is reduced by about 1/4 compared to the FC512, but the accuracies are kept about the same.

Compared with Figure 20, the accuracy decreases in case of the deeper network C14, C15 and C16. Additionally, compared with the single eye scheme and double eyes scheme, the accuracy is slightly improved due to the network being deeper.

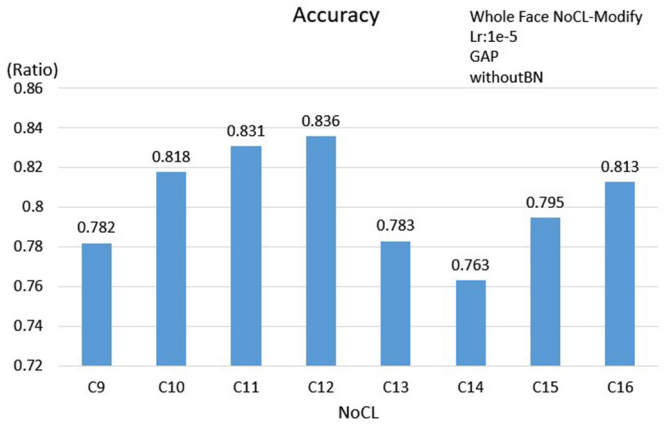

Figure 24. The accuracy of the face scheme for Experiment 3-4.

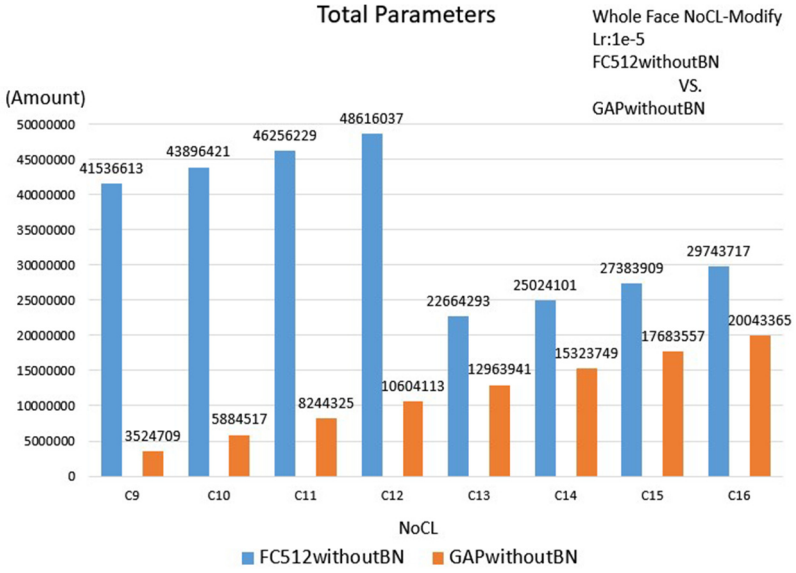

Figure 25. The number of parameters of the face scheme for replacing FC.

Experiment 3-5 evaluates the effect for different NoCL, and other parameters are set as Lr to 1×10^{-5}, with GAP and BN. The results are shown in Figure 26. The C12 has the best accuracy rate of 0.833, and the execution time for a gaze estimating is 10.776 ms. Compared to Experiment 3-4 without BN, the accuracy increases by 0.05, as shown in Figure 27.

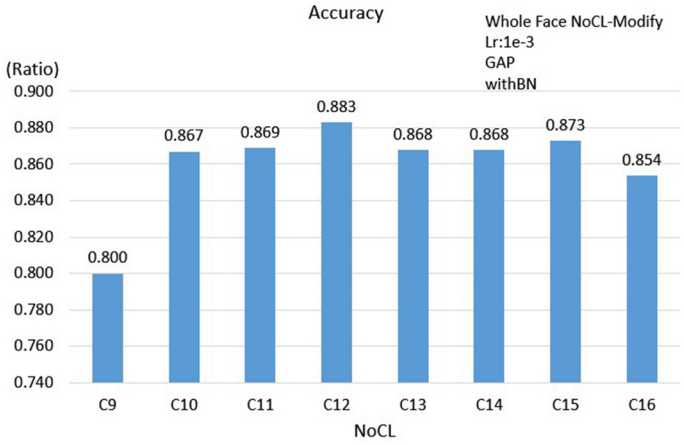

Figure 26. The accuracy of the face scheme for Experiment 3-5.

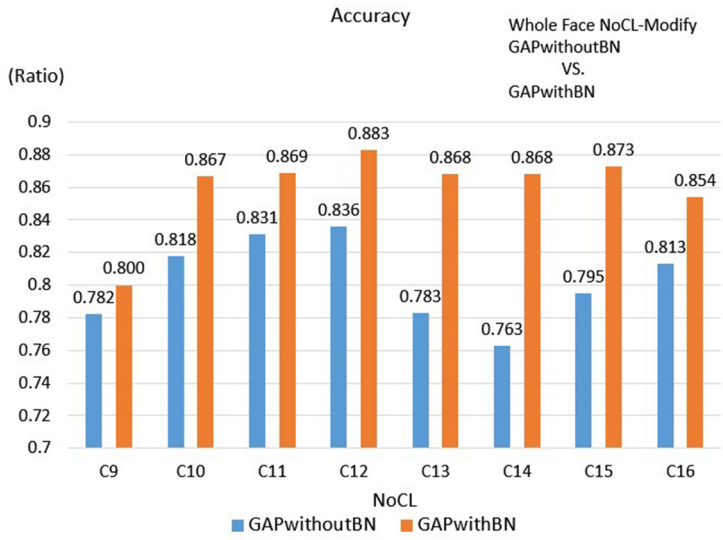

Figure 27. Comparing the accuracy of Experiment 3-5 with Experiment 3-4.

Finally, the comparison of four experiments with good performance is of the face scheme shown in Table 3. When GAP replaces the fully connected layer, the parameters reduced to 1/2 of the original. The accuracy increases to 0.883. The model architecture of the best performance includes 12 convolutional layers, BN and GAP, and the learning rate is set to 1×10^{-3}.

Table 3. Comprehensive evaluation of the face scheme.

	FC512withoutBN	FC512withBN	GAPwithoutBN	GAPwithBN
NoCL	C13	C13	C12	C12
Learning Rate	1×10^{-5}	1×10^{-4}	1×10^{-5}	1×10^{-3}
Accuracy	0.862	0.866	0.836	0.883
Loss	0.608	0.571	0.603	0.613
Precision	0.861	0.864	0.838	0.885
Recall	0.865	0.865	0.843	0.885
F1-Score	0.862	0.864	0.840	0.885
Parameters	22664293	22684261	10604113	10617957

Table 4 summarizes the results for the 3 proposed schemes. According to the single eye scheme, comparing FC512 without BN and FC512 with BN, there is no improvement in accuracy with BN. However, when using GAP, the result shows improvement in accuracy. Based on the evaluation results of the double eyes scheme, the accuracy of using BN is improved. In the face scheme, it is shown that the accuracy is significantly improved when using GAP and BN at the mean time. Overall, the highest accuracy is obtained by using GAP and BN together. In addition, the accuracy of the face scheme is the highest; however, the accuracy of the single eye scheme is the lowest.

Table 4. Comparing the accuracy of all schemes.

Scheme	FC512withoutBN	FC512withBN	GAPwithoutBN	GAPwithBN
Single Eye	0.704	0.689	0.699	0.740
Double Eyes	0.801	0.827	0.783	0.839
Face	0.862	0.866	0.836	0.883

If the numbers of input features are different, then the corresponding NoCL will be different when different schemes respectively perform the best. In this case, the face scheme obtains the maximum amount of parameters and the single eye scheme obtains the minimum amount of parameters. Table 5 presents the numbers of the parameters of the 3 proposed schemes. Overall, using GAP, the parameters have been significantly reduced.

Table 5. The numbers of the parameters of all schemes.

Parameters	FC512withoutBN	FC512withBN	GAPwithoutBN	GAPwithBN
Single Eye	3524709	4714597	2335077	2340709
Double Eyes	6933605	9307237	1744997	2340709
Face	22664293	22684261	10604113	10617957

To prove the effectiveness of our proposed face scheme with BN and GAP, we have compared it with the existing ROI and ERT methods [16] for 7 class case including UpLeft, UpRight, Left, Centre, Right, DownLeft and DownRight. A detection accuracy comparison of our model with the existing methods of George and Routray is summarized in Table 6. It indicates that our proposed method achieves better prediction accuracy of 6% and higher.

Table 6. Comparison of prediction accuracy of various methods.

	Proposed	George (ROI)	George (ERT)
Accuracy	93.3	81.37	86.81

According to Figure 28, it shows confusion matrix for the proposed face scheme. For example, for the Center class of our proposal, the incorrect predictions occur in the Left, Right and DownLeft classes. For the UpRight class in our proposal, the incorrect prediction occurs in the Right class. We observed the incorrect predictions and found that the incorrect estimated class is usually near the correct class. When the gaze estimation is applied to discover the object of interest, based on our previous work [18], the influence caused by incorrect prediction will be reduced by using the classification method.

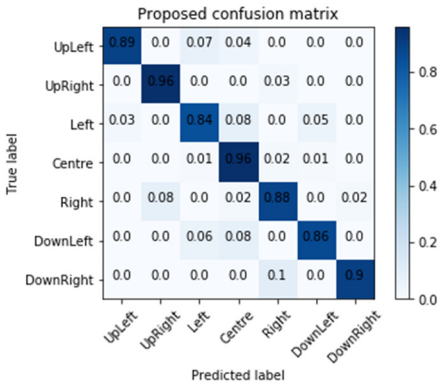

Figure 28. Confusion matrix for the proposed face scheme.

4. Conclusions

The evaluations for diversity parameters are performed by adjusting the numbers of NoCL and settings of BN as well as the GAP instead of the fully connected layer. A novel method is proposed to build a training dataset as the participant watches videos, because this is closer to the viewer's visual behavior. The most real and natural data of users can be obtained; the participants can swing their head freely without too many restrictions in the data collecting procedure. We propose three

schemes, namely, the single eye image, double eyes image and facial image to evaluate the efficiency and computing complexity under different network architectures. Generally, the input image of an eye tracking system mostly is the eye or face of the small size image with relatively few features. Regarding the efficiency of BN and GAP, this paper completed the evaluation of the 3 schemes. Based on Table 4, the results show that BN and GAP are helpful in overcoming the problem to train models and in reducing the network complexity; however, the accuracy does not necessarily show a significant improvement. It is shown that the accuracy is significantly improved when using GAP and BN at the mean time. Overall, the face scheme has the highest accuracy of 0.883 when BN and GAP are used at the mean time. Additionally, comparing to the FC512 case, the number of parameters is reduced less than 50% and the accuracy is improved by about 2%.

Since the numbers of input features are different and the corresponding NoCL at the best performance is different for the 3 proposed schemes, the number of parameters will be different. Based on Table 5, presenting the numbers of the parameters of the 3 proposed schemes, the face scheme obtains the maximum number of parameters and the single eye scheme obtains the minimum number of parameters. Meanwhile, using GAP, the parameters were significantly reduced. Therefore, the execution time for applying GAP can be reduced; in our case, there are 4.466 ms, 5.071 ms and 10.776 ms respectively for the single eye scheme, double eyes scheme and face scheme.

According to Figure 28, we observed and found that the incorrect estimated class is usually near the correct class. When the gaze estimation is applied to discover the object of interest, based on our previous work [18], the influence caused by incorrect prediction will be reduced by using the classification method. In the future, we plan to work on adaptively adjusting the block size based on the content and gaze distribution to reduce incorrect prediction.

Author Contributions: Conceptualization, H.-H.C. and B.-J.H.; Data curation, H.-H.C., B.-J.H. and P.-T.L.; Investigation, B.-J.H.; Methodology, H.-H.C. and B.-J.H.; Software, P.-T.L.; Validation, H.-H.C., B.-J.H. and J.-S.W.; Writing—original draft, H.-H.C. and P.-T.L.; Writing—review & editing, H.-H.C., B.-J.H. and J.-S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hicks, C.M.; Petrosoniak, A. Peak performance: Simulation and the nature of expertise in emergency medicine. *Can. J. Emerg. Med.* **2019**, *21*, 9–10. [CrossRef] [PubMed]
2. Laddi, A.; Prakash, N.R. Eye gaze tracking based directional control interface for interactive applications. *Multimed. Tools Appl.* **2019**, *78*, 31215–31230. [CrossRef]
3. Paul, I.J.L.; Sasirekha, S.; Maheswari, S.U.; Ajith, K.A.M.; Arjun, S.M.; Kumar, S.A. Eye gaze tracking-based adaptive e-learning for enhancing teaching and learning in virtual classrooms. In *Information and Communication Technology for Competitive Strategies*; Springer: Singapore, 2019; pp. 165–176.
4. Hansen, D.W.; Ji, Q. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 478–500. [CrossRef] [PubMed]
5. Duchowski, A. *Eye Tracking Methodology: Theory and Practice*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007.
6. Blignaut, P.; Wium, D. The effect of mapping function on the accuracy of a video-based eye tracker. In Proceedings of the 2013 Conference on Eye Tracking South Africa; ACM: New York, NY, USA, 2013; pp. 39–46.
7. Zhu, Z.; Ji, Q. Novel Eye Gaze Tracking Techniques under Natural Head Movement. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 2246–2260. [PubMed]
8. Zhou, X.; Cai, H.; Shao, Z.; Yu, H.; Liu, H. 3D eye model-based gaze estimation from a depth sensor. In Proceedings of the 2016 IEEE International Conference on Robotics and Biomimetics (ROBIO), Qingdao, China, 3–7 December 2016; pp. 369–374.

9. Anuradha, A.; Corcoran, P. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access* **2017**, *5*, 16495–16519.
10. Wu, Y.L.; Yeh, C.T.; Hung, W.C.; Tang, C.Y. Gaze direction estimation using support vector machine with active appearance model. *Multimed. Tools Appl.* **2012**, *70*, 2037–2062. [CrossRef]
11. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE Computer Society: Washington, DC, USA, 2015.
12. Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; Torralba, A. Eye tracking for everyone. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2176–2184.
13. Wang, Y.; Shen, T.; Yuan, G.; Bian, J.; Fu, X. Appearance-based gaze estimation using deep features and random forest regression. *Knowl. Based Syst.* **2016**, *110*, 293–301. [CrossRef]
14. Lemley, J.; Kar, A.; Drimbarean, A.; Corcoran, P. Convolutional Neural Network Implementation for Eye-Gaze Estimation on Low-Quality Consumer Imaging Systems. *IEEE Trans. Consum. Electron.* **2019**, *65*, 179–187. [CrossRef]
15. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. It's written all over your face: Full-face appearance-based gaze estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 2299–2308.
16. George, A.; Routray, A. Real-time eye gaze direction classification using convolutional neural network. In Proceedings of the 2016 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 12–15 June 2016.
17. Zhang, C.; Yao, R.; Cai, J. Efficient eye typing with 9-direction gaze estimation. *Multimed. Tools Appl.* **2017**, *77*, 19679–19696. [CrossRef]
18. Kao, C.W.; Chen, H.H.; Wu, S.H.; Hwang, B.J.; Fan, K.C. Cluster based gaze estimation and data visualization supporting diverse environments. In Proceedings of the International Conference on Watermarking and Image Processing (ICWIP 2017), Paris, France, 6–8 September 2017.
19. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
20. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2014**, arXiv:1312.4400.
21. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; Computer Vision Foundation: Washington, DC, USA, 2015.
22. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556.
23. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
24. Klare, B.F.; Klein, B.; Taborsky, E.; Blanton, A.; Cheney, J.; Allen, K.; Grother, P.; Mah, A.; Jain, A.K. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1931–1939.