IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Automated Diabetic Retinopathy Detection Based on Binocular Siamese-like Convolutional Neural Network

XIANGLONG ZENG[1], HAIQUAN CHEN[1], YUAN LUO[2] AND WENBIN YE [2], (MEMBER, IEEE)

[1]School of Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China
[2]School of Electronic Science and Technology, Shenzhen University, Shenzhen, 518060, China

Corresponding author: Wenbin Ye (e-mail: Yewenbin@szu.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant 61601301, in part by the Fundamental Research Foundation of Shenzhen under Grant JJCYJ20170302151123005.

ABSTRACT Diabetic retinopathy (DR) is an important causes of blindness worldwide. However, DR is hard to detected in early stages and the diagnostic procedure can be time-consuming even for experienced experts. Therefore, a computer-aided diagnosis method based on deep learning algorithms is proposed to automatedly diagnose the referable deiabetic retinopathy (RDR) by classifying color retinal fundus photographs into two grades. In this work, A novel convolutional neural network model with Siamese-like architecture is trained with transfer learning technique. Different from previous works, the proposed model accepts binocular fundus images as inputs and learns their correlation to help making prediction. In the case with a training set of only 28104 images and a test set of 7024 images, an area under the receiver operating curve (AUC) of 0.951 is obtained by the proposed binocular model, which is 0.011 higher than that obtained by existing monocular model. To further verify the effectiveness of the binocular design, a binocular model for five-class DR detection is also trained and evaluated on a 10% validation set. The result shows that it achieves a kappa score of 0.829 which is higher than that of existing non-emsemble model.

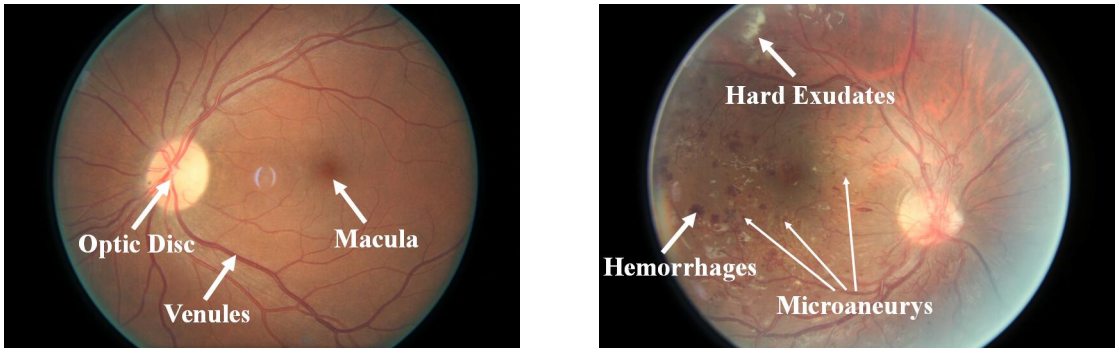
INDEX TERMS biomedical imaging processing; diabetic retinopathy; fundus photograph; convolutional neural network; deep learning; Siamese-like network

I. INTRODUCTION

DIABETIC retinopathy (DR) is a common complication of diabetes associated with retinal vascular damage caused by long-standing diabetes mellitus [1]. According to the research in [2], DR has become one of the important causes of blindness and vision impairment worldwide, since 0.4 million cases of blindness and 2.6 million cases of severe vision impairment globally can be attributed to it in 2015. In fact, the impairment of DR to vision can be controlled or averted if it is detected and treated in time. However, many patients miss the best time for treatment since there are few signs or symptoms at the early stage of DR. Furthermore, the diagnosis of DR mostly depends on the observation and evaluation to fundus photographs (see Fig.1) of which procedure can be time-consuming even for experienced experts. Therefore, computer-aided automated diagnosis approaches have great potential in clinical to accurately detect DR in a short time, which can further help to improve the screening rate of DR and reduce the number of blindness.

Multiple automated diagnosis systems have been developed over the last decade. Since human experts usually focus on some typical lesions associated with DR such as microaneurysms, hemorrhages and hard exudates (see Fig.1) when evaluating fundus photographs, many works paid attention to automatedly detect and segment these lesions or calculate some numerical indexes [3]. Shahin et al. [4] developed a system to automatedly classify retinal fundus images into those with or without proliferate diabetes retinopathy. They adopted morphological processing to extract pathological features such as blood vessels area and exudates area as well as two indexes including entropy and homogeneity. These features are fed into a shallow neural network and a sensitivity of 88% and a specificity of 100% are obtained. Jaafar et al. [5] proposed an automated algorithm, which mainly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2903171, IEEE Access



(a) (b)

FIGURE 1. Typical fundus photographs. (a) Photograph of healthy fundus, showing the normal optic disc, venules and macula. (b) Photograph of fundus with severe DR, in which three common lesions (hemorrhages, microaneurysms and hard exudates) are pointed out.

consist of two part: the top-down segmentation to segment the exudates legion and a polar coordinate system centred at the fovea to grade the severity of hard exudates. Based on a small dataset of 236 fundus images, it reaches a sensitivity of 93.2%. Casanova et al. [6] introduce an algorithm of random forest to discriminate people with or without DR with the accuracy of more than 90% and assess the DR risk based on graded fundus photographs and systemic data. However, the approach depends on clinical relevant variables such as microaneurysms count and abnormality in the graded retinal images labeled by human experts.

Most of works above either rely on the variables manually measured by experts or put much effort into extracting handcrafted features with image processing approaches which bring extra complexity and unstability. Thus, the deep learning method with the ability to learn significant features directly from the fundus photography has aroused the attention of researchers in recent years. Quellec et al. [7] proposed a system to detect referable DR by employing a deep convolutional neural network (CNN) and automatedly segment DR lesions by creating heatmaps of the convolutional layer which shows the potential to discover new biomarkers in images. They adopted the CNN structure with an ensemble learning method from the solution ranked second in the Kaggle Diabetic Retinopathy competition [8] and obtained a good detection result with the area Az under the Free-response receiver operating curve (FROC) of 0.954 in the Kaggle dataset. Gulshan et al. [9] adopted a deep CNN model named Inception V3 to detect referable diabetic retinopathy (RDR) based on a development dataset which contains more than 128 thousand fundus images. Benefited from the large training data and well sifted expert grading to the fundus images, the work achieved an impressing performance with the area under the receiver operating curve (AUC) of 0.991/0.990 and sensitivity of 97.5%/96.1% on two different test sets respectively. Rishab et al. [10] proposed a method that combines deep CNN with traditional machine learning algorithm. In their work, fundus images are fed into a residual network after pre-processing, and then the characterization of images obtained from the last pooling layer of network, appended with several metadata variables, is sent into an decision tree classifier to differentiate between healthy fundi and fundi with DR. As a result, the method achieves an AUC of 0.94 with sensitivity of 0.93 and specificity of 0.87 on a test set obtained from public.

In this paper, a deep learning based method which is inspired by the diagnostic process of human ophthalmologists is proposed to automatically classify the fundus photographs into 2 types – with or without RDR. In this work, instead of adopting fundus images of single eye as input like most previous works did, we built a novel Siamese-like CNN model with weight-sharing layers based on Inception V3, which is able to accept fundus images of both eyes as inputs and outputs the classification result of each eye at the same time. To be better adapted to the model, those binocular fundus images have been paired and pre-processed correspondingly before being fed into the network. The proposed binocular is compared with the monocular model which directly transferred from original Inception V3 with reference to [9]. The classification results of both models are evaluated with the AUC score and it shows that proposed binocular model outperforms the monocular model by a margin. Besides, a binocular model for the five-class DR detection task is also trained and evaluated to further prove the effectiveness of the binocular design. The result shows that, on a 10% validation set, the binocular model achieves a kappa score of 0.829 which is higher than that of existing non-emsemble model. Finally, the comparison between confusion matrices obtained through models with paired and unpaired inputs is performed and it demonstrates that the binocular architecture does improve the classification performance.

The remaining of this paper is organized as follows. The implementation of our method and the detailed architecture of the binocular Siamese-like network are described in Section II. The experimental results and evaluation of the model performance are showed in Section III. In the end, the discussion and conclusion are given in section IV.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2903171, IEEE Access

II. PROPOSED METHOD

For a deep learning model, the most important parts that should be focused on are data set, network architecture and training method. Before being used to train our model, fundus images data set obtained from public resources is pre-processed and augmented, and the detail is discussed in Section II.A. The architecture of our network and design philosophy are described in Section II.B. The training detail of the model is laid out in Section II.C.

A. IMAGE PRE-PROCESSING AND AUGMENTATION

Data set is a primary and significant part that need to be dealt with for a deep learning application. The fundus photographs in our data set have large variation, such as discrepant brightness or resolution, since most of them are obtained with different equipment in different environment. Thus, in order to standardize these images, reduce redundant information and environmental artifacts, several pre-processing methods are applied to the fundus images:

1) Scale down these images according to their original aspect ratio to keep the short side of images 299 pixels. Then clip the long side of images and retain the center 299 pixels. This step is to unify the size of images into 299x299 pixels so that it can compared with the monocular mothod proposed in [9].
2) Subtract each pixel value of images by the weighted means of its surrounding pixel values, and add it by 50% grayscale. This operation is similar to the 'high pass' processing in the PhotoShop software, which makes the blood vessels as well as the lesion areas in fundus images more explicit. Then, the fundus area in images will be clipped to 95% of the original size by covering a mask with a transparent circle on the center in order to remove the boundary effects arising from the last operation. The processing method in this step is refer to the algorithm proposed by [11].
3) Convert the pixel values of images from [0, 255] to [-1, 1] before feeding images to the network. This step is to normalize the input data, help it better propagate through the network and avoid the negative effect from ill-conditioned values.

The other problem of the image data set is that it's too small for a deep learning model to solve a medical image recognition problem with high accuracy. Therefore, besides of the pre-processing steps, multiple image augmentation steps are further imposed on the data set in order to improve the generalization performance of the proposed model. But before the augmentation, it should be noted that the original fundus images contains many physiological information of patients. For example, although some fundus images are inverted due to different imaging modes of fundus cameras, one can still identify a specific image is obtained from left eye or right eye since the connection line through the macula and optic nerve is always with negative slope for the left eye and positive slope for the right eye. The physiological information and relative position relation like this should be preserved for our binocular model. Moreover, the augmentation is actually performed between the step 1) and step 2) of pre-processing and all the steps are symmetrically and synchronously applied to binocular fundus images from certain patients. The detailed augmentation steps are as below:

1) Randomly flip the images of left eye and right eye horizontally. Since human eyes are structurally mirror-symmetric, the binocular images can be flipped horizontally and then exchanged before being input to the network.
2) Randomly perform geometric transformation on images, including randomly inverting images, cropping images by 0%-5% of their height of width, scaling to 90%-110% of their size, translating by -5 to +5 pixels, rotating from -30 to 10 degrees and shearing from -10 to 10 degrees.
3) Randomly change the brightness and contrast of images, including adding or subtracting the image value by 10, change the brightness to 85%-115%, decreasing or improving the contrast in the range of 85%-115%.

All the steps and substeps of augmentation listed above are performed during the training process with a probability of 50%.

B. NETWORK ARCHITECTURE AND IMPLEMENTATION

The deep learning model proposed in this paper is a novel convolutional neural network with Siamese-like architecture of which block diagram is shown in Fig.2. Basically, the model accepts two fundus images corresponding to the left eye and right eye as inputs and then transmits them into the Siamese-like blocks. The information from two eyes is gathered into the fully-connected layer and finally the model will output the diagnosis result of each eye respectively, i.e., with or without referable DR. To a large extent, the pipeline of our model is inspired by the clinical diagnosis process of DR in real life. The detail of the model block, as well as the explanation of the network architecture is laid out in subsections below.

1) Inception V3

Inception V3 [12] is a well known deep CNN model. The basic architecture of Inception V3 is shown in Fig.3. The outstanding performance of Inception V3 is benefited by several network connection techniques, such as adopting batch normalization, using MLPconv layers to replace linear convolutional layers and factorizing convolutions with large kennel size [12] [13]–[15]. With these techniques, the number of the network parameters as well as the computational complexity are reduced significantly, and thus the network can be built much deeper and get stronger non-linear expressive ability than that of the conventional CNN models.

By adopting the transfer learning method, Inception V3 can be customized to perform different image classification tasks. For example, it was modified to make several binary

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2903171, IEEE Access

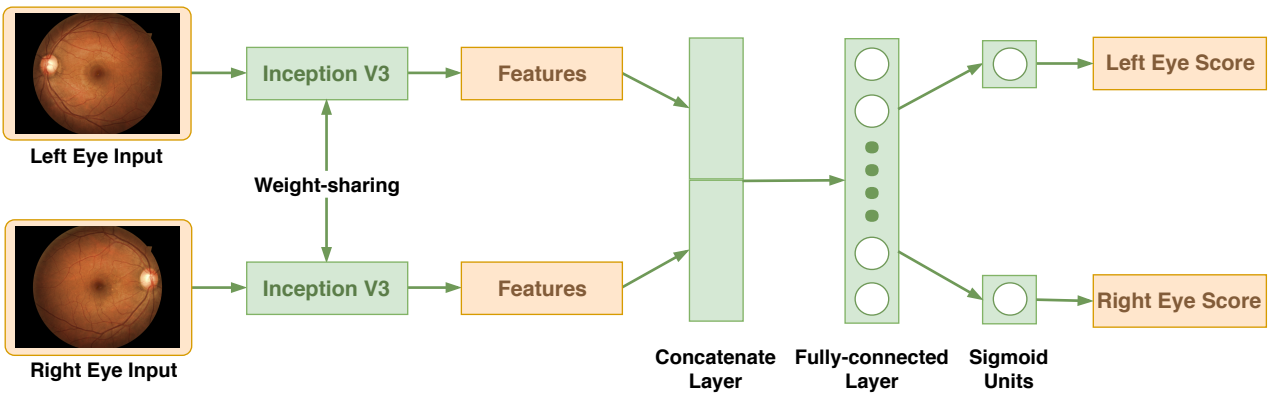


FIGURE 2. **Block Diagram of the proposed Siamese-like binary classification convolutional neural network.**

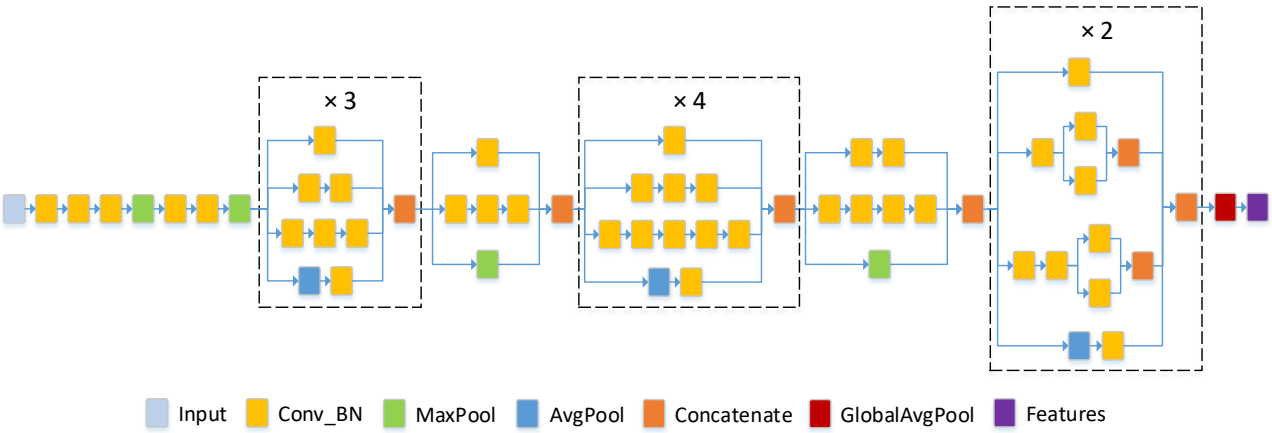


FIGURE 3. **Block Diagram of Inception V3 architecture without the last layer. Conv_BN represents convolutional layer with batch normalization, MaxPool represents max pooling layer and AvgPool represents average pooling layer. "× t" represents that the structures in dashed blocks are repeated for t times.**

predictions by replacing the final layer with customized layers to classify different stages of DR [9]. In this work, Inception V3 is embedded into the Siamese-like blocks of the network after removing the final layer in order to extract high-dimensional features of fundus images. Besides, the original rectified linear unit (ReLU) activation function used in Inception V3 is replaced by Leaky ReLU, since the latter one has a better characteristic of gradient propagation for large CNN models. The leaky rate is set to be 0.1.

2) Siamese-like Network Structure

In most cases, when patients take fundus examinations in the hospital, both of their eyes will be photographed. And rather than just watching the photograph of a single eye, ophthalmologists will put fundus photographs of both eyes together and make a diagnosis to retinal disease by referring to these two photographs and comparing them with each other since the physiological and pathological conditions of one eye has important guiding significance for the diagnosis to the other eye. For example, if someone's left eye has obvious symptoms of severe DR, then there will be a strong indication that he/she have suffered from diabetes mellitus for a long time and the other eye is very likely to be with DR. Thus, it give us a inspiration to design a binocular network with Siamese-like structure, which is able to accept binocular fundus images synchronously, gather their features and make the prediction of each eye.

In fact, the actual Siamese network, as shown in Fig.4, is built to compare the similarity of two inputs in the feature space [16]. First, two inputs, X_1 and X_2, are sent into two network blocks which have same architecture G and shared weights w. Then, feature vectors of two inputs, $G_w(X_1)$ and $G_w(X_2)$, are obtained from the last layer and their distance (e.g., Euclidean distance), $<G_w(X_1), G_w(X_2)>$, will be calculated. If the distance is small, it can be inferred that these two inputs are of high similarity.

But the similarity of fundus images in the last step is not going to be estimated, instead, the correlation between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2903171, IEEE Access

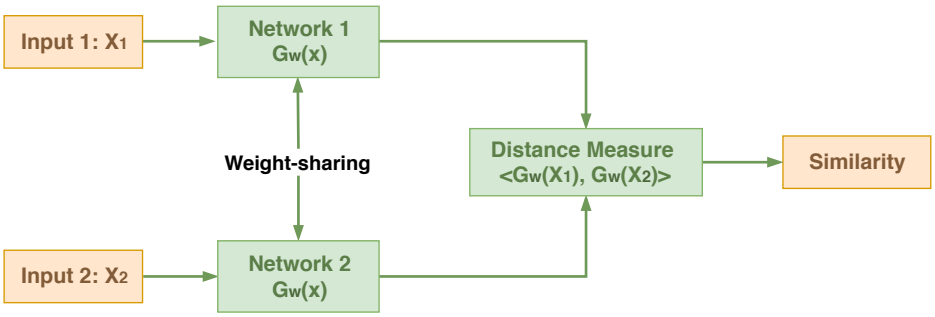


FIGURE 4. Block Diagram of the Siamese network. X_1 and X_2 are two inputs of the network, $G_w(X_1)$ and $G_w(X_2)$ represent their corresponding feature vectors and <$G_w(X_1), G_w(X_2)$> represents the distance between two vectors.

binocular images is going to be utilized to assist the prediction. So, in our work, the distance calculation part is taken away as shown in Fig.2. Two weight-sharing blocks, with the same Inception V3 architecture, are adopted to extract high-dimensional features of binocular fundus images (i.e., $G_w(X_1)$ and $G_w(X_2)$) since ophthalmologists will similarly point out lesions in both eyes based on the same criteria in clinic. Then, these features will be concatenated together and connected with a fully-connected layer in order to integrate the information of both eyes and let the network to decide which features are important and have reference value when making the prediction of each eye. The last layer of the entire network contains two classification units with sigmoid activation function corresponding to predictions of two eyes respectively. The prediction result of each eye will be a continuous number between 0 to 1 representing the probability that the eye is with referable DR.

C. TRAINING METHOD

Transfer learning method is a widely used training method of neural network [17] which has been mentioned in Section II.A.1 and it is also adopted to make our model trained more efficiently. By loading the weights of Inception V3 blocks pre-trained on ImageNet data set, the model will has a better weights initialization before starting the gradient optimization. Moreover, considering the huge difference between the fundus images data set and ImageNet data set, none of layers in weight-sharing Inception V3 blocks are frozen. In other words, every layer in the blocks is trainable.

Beside, the selection of optimizer is important when training a deep learning model. A good optimizer can remarkably speed up the training process, avoid the bad local optima and give us a better training result. The optimizer used in this work is called Adam [18]. It is a widely-used method with adaptive learning rate which has been proven to be effective and practical on deep learning optimization.

To prevent the network from over-fitting and improve the prediction performance on the test set, "dropout" [19] is imposed on the fully-connected layer. This technique will randomly drop out units in the network during training process in order to reducing the co-adaptation on the training set and enhance the generalization ability of network.

Other techniques, such as "early stop" and "fine tune", are also adopted in our work. They are useful to shorten time for training and tuning, helping us to obtain models with better performance.

III. RESULTS

A. DATA SET

The image dataset used in our work, which is obtained from the website of Kaggle diabetic retinopathy competition [8] provided by EyePACS, contains 35126 high resolution fundus photographs taken under a range of imaging conditions. These fundus photographs have been labeled by a trained clinician with a scale of 0 to 4 based on the severity of DR. Table 1 shows the 5 classes of DR as well as their respective proportion. According to the International Clinical Diabetic Retinopathy Scale [20], RDR is defined as the presence of moderate and worse DR and/or referable diabetic macular edema. Thus, images with labels of 0 and 1 are classified as "without RDR" and relabeled with 0, images with labels of 2, 3 and 4 are classified as "with RDR" and relabeled with 1. The distribution of relabeled data set is shown in Table 2.

TABLE 1. The Distribution of Original Data

Label	Class	Number	Percentage
0	No DR	25810	73.5%
1	Mild	2443	6.9%
2	Moderate	5292	15.1%
3	Severe	873	2.5%
4	Proliferative DR	708	2.0%

TABLE 2. The Distribution of Relabeled Data

Label	Class	Number	Percentage
0	No RDR	28253	80.4%
1	RDR	6873	19.6%

It's obvious that the data set has the problem of class-imbalance. In order to mitigate this problem, binocular fun-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2903171, IEEE Access

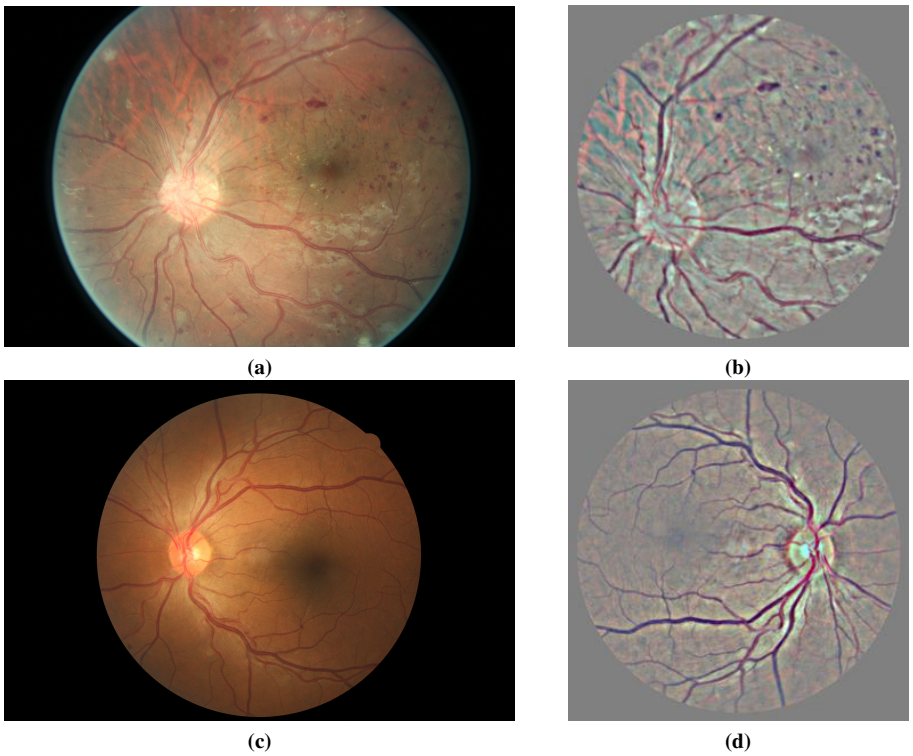


FIGURE 5. Examples of original fundus images and the corresponding processed images. (a) and (c) Original images. (b) and (d) Processed images.

dus images are group into four bunches in pairs: both eyes are with RDR, both eyes are without RDR, only the left eye is with RDR and only the right eye is with RDR. Then 80% images in each bunch are stored into the training set jointly while the remaining 20% of each bunch will be used as test set, which ensures the proportion of images with different labels is same in both training set and test set.

Besides, it is noteworthy that these images are taken by different types of cameras in different environment. Some of the images are inverted while some are incomplete due to the variant microscope imaging procedure. Also, noise in both images and labels is unavoidable in the raw data set, which puts forward a high demand to the robustness of our classification system.

B. PRE-PROCESSING AND AUGMENTATION

Two examples of the raw fundus photographs, as well as the corresponding processed images after pre-processing and augmentation, are showed in Fig.5. It can be seen that the fundus region in Fig. 5(c) is of well integrity and the imaging quality is high, but the fundus region in Fig. 5(a) is incomplete and there are some illumination artifacts around its edges. After pre-processing, the marginal region of fundus in Fig.5(a) is clipped, which removes the artifacts and makes the remaining fundus region a complete circle as shown in Fig.5(b). Besides, the fundus area in Fig.5(b) has been rotated by a small angle and Fig.5(b) is flipped compared with Fig.5(a). The rotation and flipping, resulted from augmentation, are randomly performed on fundus images. And from both Fig.5(b) and Fig. 5(d), it can be found that fundus regions are standardized to be circular areas with a same size, containing the optic disc, macula and main vessels. Furthermore, the background color of retina is faded, while the venules, hard exudates and hemorrhages are emphasized.

C. EVALUATION

The proposed binocular model for RDR detection, as well as a similar model for the original five stage DR classification task are trained on a single server with NVIDIA GeForce GTX1080TI graphics cards. They are evaluated with different metrics in the following two subsections. Note that during the training and evaluation, images of twos eyes from corresponding patients are paired before being sent into the networks, since the proposed binocular models are designed to work in the case that fundus images of both eyes are available.

1) Model for RDR Detection

The proposed model for RDR Detection is trained on the server for more than 24 hours. For comparison, a monocular Inception V3 model is also trained with transfer learning method according to the description in [9]. The best versions of these two models are selected after multiple times of training and their performance is evaluated by the AUC score, which represents the area under the receiver operating curve (ROC). In most case, the output of binary classification model is a probability value from 0 to 1. If different thresholds are selected, there will be different classification results,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2903171, IEEE Access

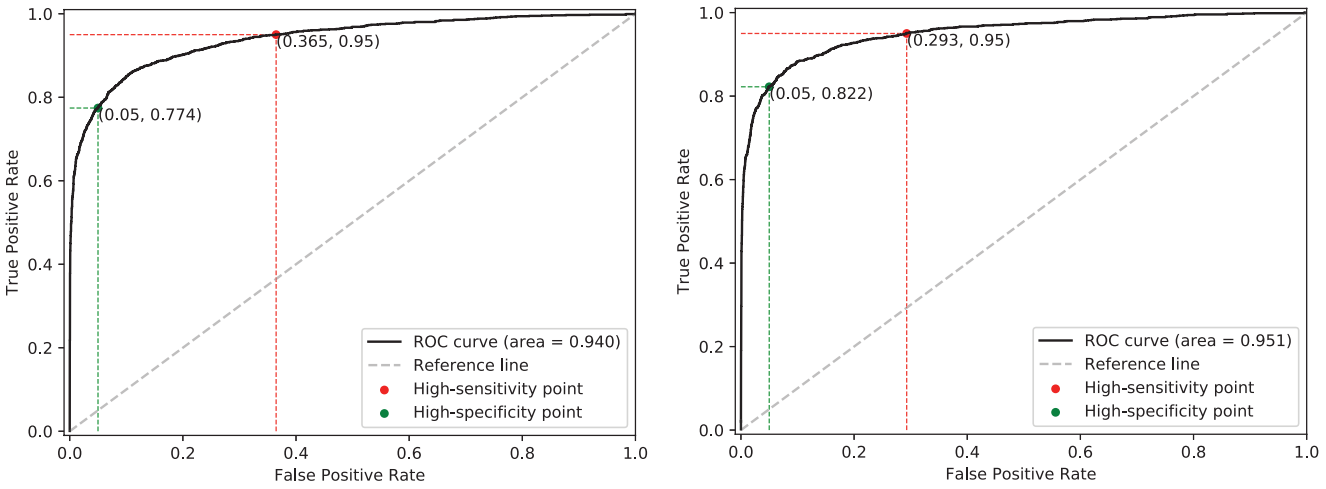


FIGURE 6. ROC of two models. Note that true positive rate equals to sensitivity and false positive rate numerically equals to 1 - specificity. The gray dotted line is the reference line which represents the ROC with an AUC of 0.5 resulting from random guessing. (a) ROC of monocular Inception V3 model, with a sensitivity of 77.4% on the high specificity operating point and a specificity of 63.5% on the high sensitivity operating point. (b) ROC of proposed binocular model, with a sensitivity of 82.2% on the high specificity operating point and a specificity of 70.7% on the high sensitivity operating point.

TABLE 3. The Evaluation Result of Proposed Binocular Model and Monocular Model

Model	AUC	Sensitivity (Specificity=95%)	Specificity (Sensitivity=95%)
Monocular model	0.940	77.4%	63.5%
Binocular model	0.951	82.2%	70.7%

bringing the model different sensitivities and specificities. (The definitions of sensitivity and specificity are given in Appendix A.) Thus, the ROC can be generated by plotting sensitivity (i.e., true positive rate) vs 1 - specificity (i.e., false positive rate) since each threshold is corresponds to a pair of sensitivity and specificity. The better the classification model is, the closer the curve is to the top left corner and the more approaching the AUC score is to 1.

The ROC of proposed model and monocular model is shown in Fig.6. As shown in Fig.6(a) and Fig.6(b) respectively, the AUC of monocular model is 0.940 and the AUC of proposed model is 0.951. To compare the performance of two models in detail, two operating points on the ROC are selected with reference to [9]. The first operating point is a high-sensitivity point of which sensitivity is fixed at 95.0% and the second operating point is a high-specificity point of which specificity is fixed at 95.0%. As annotated on the Fig.6(a) and Fig.6(b), for the high-specificity operating points, the sensitivity of proposed model is 0.80 while the sensitivity of monocular model is 77.4%. For the high-sensitivity operating points, the specificity of proposed model is 82.2% while the specificity of monocular model is 70.7%. For more intuitive comparison, the result is also laid out in Table 3. It demonstrates that proposed binocular model achieves higher performance than monocular Inception V3 model on both operating points, which means binocular model has greater potential in clinical application.

2) Model for Five Stages DR Classification

In order to further verify the effectiveness of binocular architecture, a slightly altered binocular network is applied to the original five-class classification task of Kaggle competition. Specifically, the binary classification units in the last layer of our network is simply replaced with 5 units and the input image size is changed into 512x512 pixels so that the model can be compared with the existing model developed by team o_O in the Kaggle competition. The data partition method is similar to that used in RDR detection, except that training set and validation set are split with 9:1 ratio and their labels are is restored to five classes. Quadratic weighted kappa score is adopted to evaluate performance of the five-class classification model. It is able to describe the agreement between predicted labels and true labels more effectively, since it cost-sensitively penalizes the wrong predictions in terms of the degree of classification error.

The kappa score of the proposed binocular model, monocular model and model form team o_O are shown in Table 4. It should be noted that the team o_O's kappa score shown in the Kaggle's leaderboard is obtained through blending and ensembling method. Thus, for a more fair comparison, team o_O's kappa score calculated on 10% validation set without using blending and ensembling is obtained from [21] and listed on the table. One can see that the proposed binocular model achieves highest kappa score, which is 0.1 higher than that of team o_O's model and significantly higher than that of the monocular model.

In addiction, Fig.7 (a-c) respectively show the confusion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2903171, IEEE Access

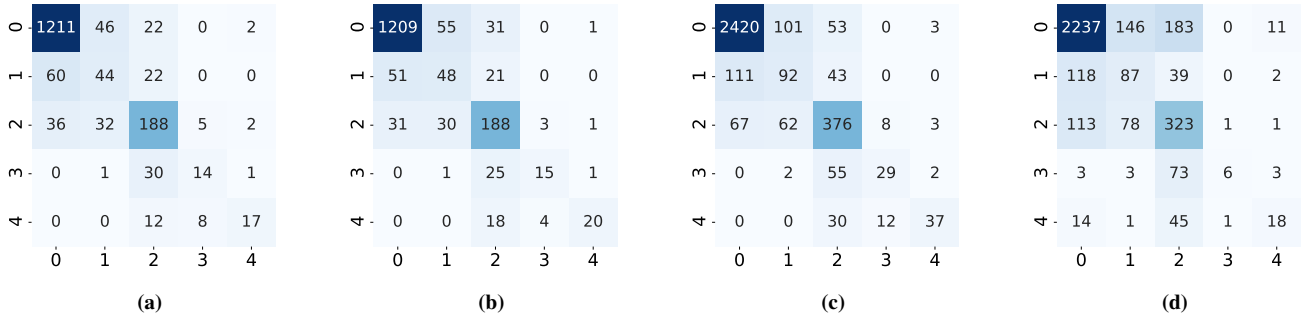


FIGURE 7. Confusion matrices obtained through binocular network, where the vertical coordinate corresponds to the labels of ground truth and the horizontal coordinate corresponds to the labels of prediction results. Suppose C{i,j} is the value in a certain grid, then it denotes the number of samples which are with true label j and predicted to be i. (a) Confusion matrix of left eyes. (c) Confusion matrix of right eyes. (b) Confusion matrix of both eyes together. (d) Confusion matrix of both eyes, but input fundus images are not paired.

TABLE 4. The Kappa Score of Different Models Based on 10% Validation Set

Model	Kappa score
Monocular model	0.808
Team o_O's model	0.800
Binocular model	0.829
Binocular model with unpaired inputs	0.620

matrices of prediction results of left eye, right eye, and both eyes together. These results are obtained based on our 10% validation set. The prediction results of the left eye and the right eye have very similar distribution patterns, indicating that the data partition method preserves the original image categories' distribution of left eyes and right eyes well. It can be found that most of the samples of class "0" and "2" are classified correctly, as the number of these samples are basically clustered on the diagonal line. However, there are relatively more samples of class "1" are misclassified to be class "0". It can be inferred that the early stages of DR are more difficult to be correctly detected and the class-imbalanced training data (where most images are labeled with "0"), to a certain extend, mislead the prediction, which finally drag down the kappa score. The classification result of images with label '3' is not satisfactory neither, probably because there is too little training data for class '3'. To more intuitively prove that the correlation between two eyes from one patient helps the prediction and improves the performance indeed, the confuse matrix, which is obtained in the case that input images are not paired (i.e., not from same patients) is shown in Fig.7. One can see that more samples deviate from the diagonal line. Serious classification errors occurred especially for samples with true labels of "3" and "4". It can be inferred that the correlation between two eyes is useful for the prediction of images with fewer training samples. Besides, the kappa score of the model with unpaired inputs, which is significantly lower than that of the model with paired inputs, has been listed on Table 4. The result indicates that the proposed binocular design is effective.

IV. CONCLUSIONS

In this paper, A novel CNN model to automatically detect RDR based on the deep learning method is developed. The proposed model has a Siamese-like architecture which accepts binocular fundus images as inputs and predicts the possibility of RDR for each eye by utilizing the physiological and pathological correlation of both eyes. The evaluation result shows that proposed binocular model achieves high performance with an AUC of 0.951 and a sensitivity of 82.2% on the high specificity operating point and a specificity of 70.7% on the high sensitivity operating point, which outperforms that of existing monocular model based on Inception V3 network. A binocular model is also trained for the original five-class DR classification task. It achieves a kappa score higher than existing work on the 10% validation set. Besides, the model with unpaired inputs is evaluated and it confirms the effectiveness of the proposed binocular design. These results also demonstrates that the proposed model has great potential to assist ophthalmologists to diagnose RDR more efficiently and improve the screening rate of RDR.

Furthermore, the proposed binocular model can be modified and applied to the auto detection of other ophthalmic diseases without much difficulty. But there is still much room for proposed model to get improvement. For example, the binocular models will have difficulties in training or testing with other dataset in which paired fundus images are unavailable. Some measures will be experimented in the following work, e.g., developing an extra model to distinguish whether two fundus images are from the same patient, and using a symmetrical image to act as the missing image of the other eye for the case that only a single fundus image is available. The other bottleneck is the small size of our data set. Predictably the performance of model can be further improved if more data are collected in the future.

.

APPENDIX A CALCULATIONS OF SENSITIVITY AND SPECIFICITY

Sensitivity and Specificity are two metrics widely used to evaluate the performance of binary classifier as well as the result of clinical examination. Sensitivity represents the rate of classifying true positive samples as positive. Specificity represents the rate of classifying true negative samples as negative. Their definition is as below:

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{FP + TN} \tag{2}$$

where TP represents the number of true positive samples in the prediction result. Analogously, FP represents the number of false positive samples, FN represents the number of false negative samples and TN represents the number of true negative samples.

APPENDIX B CALCULATION OF QUADRATIC WEIGHTED KAPPA

The quadratic weighted kappa is an metric to measure the agreement between two ratings which widely adopted in the multi-class classification task. Its calculation formula is shown as below:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{3}$$

where O is an N x N histogram matrix in which the element O_{ij} corresponds to the number of samples that with the actual label i and received a predicted label j (i.e., the confusion matrix). E is an N x N matrix of expected ratings which equals to the outer product between the category distribution histograms of prediction result and ground truth. Both of E and O have been normalized so that they have the same sum. ω is an 5x5 matrix of weights which calculated based on the difference between ground truth and predicted result:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{4}$$

where N corresponding to the number of the classification categories, which equals to 5 in this work.

REFERENCES

[1] N. Cheung, G. Tikellis, and J. J. Wang, "Diabetic retinopathy." Ophthalmology, vol. 114, no. 11, p. 2098, 2010.

[2] S. R. Flaxman, R. R. Bourne, S. Resnikoff, P. Ackland, T. Braithwaite, M. V. Cicinelli, A. Das, J. B. Jonas, J. Keeffe, J. H. Kempen et al., "Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis," The Lancet Global Health, vol. 5, no. 12, pp. e1221–e1234, 2017.

[3] A. Ahmad, A. B. Mansoor, R. Mumtaz, M. Khan, and S. H. Mirza, "Image processing and classification in diabetic retinopathy: A review," in European Workshop on Visual Information Processing, 2015, pp. 1–6.

[4] E. M. Shahin, T. E. Taha, W. Al-Nuaimy, S. E. Rabaie, O. F. Zahran, and F. E. A. El-Samie, "Automated detection of diabetic retinopathy in blurred digital fundus images," in Computer Engineering Conference, 2013, pp. 20–25.

[5] H. F. Jaafar, A. K. Nandi, and W. Al-Nuaimy, "Automated detection and grading of hard exudates from retinal fundus images," in Signal Processing Conference, 2011 European, 2011, pp. 66–70.

[6] R. Casanova, S. Saldana, E. Y. Chew, R. P. Danis, C. M. Greven, and W. T. Ambrosius, "Application of random forests methods to diabetic retinopathy classification analyses," Plos One, vol. 9, no. 6, p. e98587, 2014.

[7] G. Quellec, K. Charria"Re, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," Medical Image Analysis, vol. 39, pp. 178–193, 2017.

[8] Kaggle, "Diabetic retinopathy detection," https://www.kaggle.com/c/diabetic-retinopathy-detection/, July 27, 2015, accessed May 7, 2018.

[9] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, and J. Cuadros, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." Jama, vol. 316, no. 22, p. 2402, 2016.

[10] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," Ophthalmology, vol. 124, no. 7, pp. 962–969, 2017.

[11] B. Graham, "Kaggle diabetic retinopathy detection competition report," https://kaggle2.blob.core.windows.net/forum-message-attachments/88655/2795/competitionreport.pdf/, August 6, 2015, accessed May 20, 2018.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[13] M. Lin, Q. Chen, and S. Yan, "Network in network," arXiv preprint arXiv:1312.4400, 2013.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, pp. 448–456, 2015.

[16] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 539–546.

[17] S. J. Pan, Q. Yang et al., "A survey on transfer learning," IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

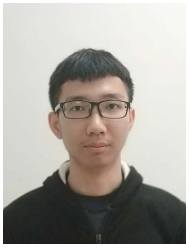
[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] E. Levels, "International clinical diabetic retinopathy disease severity scale detailed table," 2002.

[21] A. Mathis and B. Stephan, "Kaggle diabetic retinopathy detection team o_o solution," https://github.com/sveitser/kaggle_diabetic/blob/master/doc/report.pdf/, August 7, 2015, accessed January 11, 2019.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2903171, IEEE Access

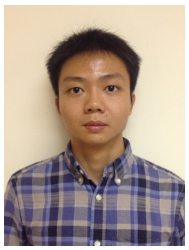


XIANGLONG ZENG received the B.E. degree in School of Electronic Science and Technology, Shenzhen University, Shenzhen, China. He currently is working toward the M.S. degree at School of Optoelectronic Engineering, Shenzhen University, Shenzhen, China.

His research interests include digital signal and image processing, machine learning and artificial intelligence.

HAIQUAN CHEN received the B.E. degree in School of Electronic Science and Technology, Shenzhen University, Shenzhen, China. He currently is working toward the M.S. degree at School of Optoelectronic Engineering, Shenzhen University, Shenzhen, China.

His research interests include radar signal analysis, digital image processing, deep learning and artificial intelligence.



YUAN LUO received the B.S. degree in Physics from Sichuan University, Chengdu, China, in 2009, and the Ph.D. degree from National University of Singapore, in 2015. Since 2016, he has been with the College of Electrical Science and Technology, Shenzhen University, where he is currently a Research Fellow.

His research interests include machine learning and bio-medical signal processing.

WEN-BIN YE (M'87) (S'12–M'14) received the B.S. degree in microelectronics from Sichuan University, Chengdu, China, in 2009, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2014. From 2014 to 2015, he was a Project Officer with Nanyang Technological University. Since 2015, he has been with the College of Electrical Science and Technology, Shenzhen University, where he is currently an Associate Professor.

His research interests include digital filter design, non-uniformly sampled data processing, Machine learning and bio-medical signal processing.