Improvement of the classification accuracy in discriminating diabetic retinopathy by multifocal electroretinogram analysis*

Xu Jin[1], Hu Guangshu[1]** and Huang Houbin[2]

(1. Department of Biomedical Engineering, Tsinghua University, Beijing 100084, China; 2. Department of Ophthalmology, General Hospital of Chinese PLA, Beijing 100853, China)

Accepted on February 13, 2007

Abstract The multifocal electroretinogram (mfERG) is a newly developed electrophysiological technique. In this paper, a classification method is proposed for early diagnosis of the diabetic retinopathy using mfERG data. MfERG records were obtained from eyes of healthy individuals and patients with diabetes at different stages. For each mfERG record, 103 local responses were extracted. Amplitude value of each point on all the mfERG local responses was looked as one potential feature to classify the experimental subjects. Feature subsets were selected from the feature space by comparing the inter-intra distance. Based on the selected feature subset, Fisher's linear classifiers were trained. And the final classification decision of the record was made by voting all the classifiers' outputs. Applying the method to classify all experimental subjects, very low error rates were achieved. Some crucial properties of the diabetic retinopathy classification method are also discussed.

Keywords: diabetic retinopathy, multifocal electroretinogram, feature selection, linear classifier.

Diabetic retinopathy (DR) is one of the major causes of blindness in modern society. It is a common micro vascular complication in patients with diabetes. As the diabetic retinopathy is a progressive disease, according to the level of severity, it could be classified as stages of DR absent, non-proliferative DR, and proliferative DR[1]. While the patients usually cannot be aware of the impairment of the disease at initial stages of diabetic retinopathy, early detection of the disease is very important.

The multifocal electroretinogram (mfERG)[2,3] is a newly developed electrophysiological technique. It allows fast deriving of electric responses from many retinal areas simultaneously. Retinal areas are illuminated by flashing hexagonal blocks on a screen placed in front of the eye. And the local ERG-like responses, each corresponding to a stimulated retinal area, can be extracted using a cross-correlation technique[4]. In recent years, mfERG was used to explore diabetic eyes with retinopathy or before clinical signs of retinopathy[5—8].

Among those researches, amplitude reduction and delay of implicit time of the mfERG responses were reported evolving with progression of the retinopathy[5—8]. But in early stages of the diabetic retinopathy, infections on mfERG response are weak and are not easy to be identified. And the recognition rate between healthy subjects, subjects without DR and subjects with early stages of DR is still low at present.

Besides the weak sign of DR shown in mfERG responses, the other reason of the low diagnosis rate of the disease is that we are short of methods to explore the abundant information that mfERG provides. Although mfERG gives many local retinal responses simultaneously, in usual processing of mfERG results[3], all these local responses are averaged into 1 to 6 traces to simplify the data analysis. And this responses averaging step obviously reduces the available information contained in mfERG result.

In this study, four groups of persons, normal ones, diabetics without apparent retinopathy, patients having non-proliferative DR and those having proliferative DR, underwent mfERG tests. A new strategy was proposed to classify the mfERG results. The strategy included definition of original features, feature selection and classifier training. Additionally, to explore the local mfERG responses, an evaluation value was computed to combine all the classifier outputs for the local responses. And the classification decision for a subject was made based on sign of the e-

* Supported by National Natural Science Foundation of China (Grant No. 30370399)

** To whom correspondence should be addressed. E-mail: hgs-dea@tsinghua.edu.cn

valuation value.

1 Experimental data acquisition

The experimental subjects to be examined by mfERG instrument included 31 eyes of healthy persons and 129 eyes of patients with non-insulin dependent diabetes mellitus (NIDDM). All the diabetic eyes also underwent ophthalmoscopic examination and fluorescein angiography. Subjects with visible media opacity or other history of ocular disease or surgery were excluded from the study.

The retinopathy was graded by experts from General Hospital of Chinese PLA, following the Early Treatment Diabetic Retinopathy Study (ETDRS)[9] protocol. Totally 58 diabetic eyes were classified as retinopathy absent (labeled as NDR in the following), 45 eyes had mild to moderate non-proliferative DR (NPDR), and 26 diabetic eyes had proliferative DR (PDR) (Table 1). The tenets of the Declaration of Helsinki were followed and the informed consent was obtained from each participant.

Table 1. Characteristics of experimental subjects

Subject	Number	Age	Gender (M/F)	Duration of diabetes(y)	Corrected visual acuity
Normal	31	45±19	14/17	/	20/20
NDR	58	61±13	34/24	7±5	20/20
NPDR	45	62±10	25/20	16±10	≥20/30
PDR	26	65±17	15/11	25±10	≥20/50

The mfERG examinations were performed on a visual evoked response imaging system (Veris Science^{TM} 5.0, Electro-Diagnostic Imaging, USA). The visual stimulation consisted of a 103-scaled-hexagons array on a 29 cm × 38 cm CRT screen, placed 40 cm in front of the tested eyes and stimulating the central retina. The luminance was 200 cd/m^2 in white hexagons and 2 cd/m^2 in black hexagons.

Each stimulus hexagon flickered at a rate of 75 Hz. The binary m-sequence which was used to control the states of the hexagons included 2^{14} steps. The raw recording time was 3 min 38 s for each mfERG record, and the duration was broken into 8 segments, 27 s long each, with a brief rest period between each segment.

Pupils were dilated maximally (≥8 mm) with eye drops. A Burian-Allen bipolar contact lens electrode was placed on the tested eye, and the other eye was patched. The retinal responses from the electrode were band pass filtered (10—300 Hz) and amplified (50000 gain).

For each mfERG record, 103 local responses were derived. Every local mfERG response corresponded to a small retinal area stimulated by a hexagonal block on the stimulus screen, named by a sequence number from 1 to 103. Each local response, covered 200 ms long, was combined with two kinds of mfERG results. The initial part waveform, 100 ms in length, was exported from software of the visual evoked response imaging system. And the waveform was treated as mfERG first order kernel[2—4] in the software. The latter 100 ms waveform, representing mfERG first slice of second order kernel[2—4] in the software, was also exported. Waveforms of all the local responses were digitized into arrays of amplitude value (Fig.1), at the sampling frequency of 1000 Hz (sampling interval was 1 ms).

2 Data analysis method

The data analysis procedure included a feature evaluation/selection step followed by a linear classification. There were 4 classes of mfERG records, labeled as Normal, NDR, NPDR and PDR in Table 1.

Amplitude value of each point on the local mfERG response was used as one potential feature to discriminate the subjects. As 200 amplitude values being obtained from each local mfERG response, the number of all available features for a mfERG recording equaled 200 features × 103 local mfERG responses, which means the original feature space has the dimension of 200 × 103. To design practical classifiers, the most informative feature subset should firstly be selected from the feature space.

2.1 Feature selection method

For the 200 original features within each local mfERG response, a feature subset having lower dimension was selected. And the criterion of inter-intra distance[10] was used to evaluate the discriminative power of the feature subset between different classes of the experimental subjects. The criterion function for the feature subset $\boldsymbol{X}_i$ is

$$\begin{aligned} J(\boldsymbol{X}_i) &= \mathrm{trace}(\boldsymbol{S}_w^{-1}\boldsymbol{S}_b) \\ \boldsymbol{X}_i &= [x_1^i, x_2^i, \cdots, x_d^i] \\ i &= 1, 2, \cdots, 103 \end{aligned} \tag{1}$$

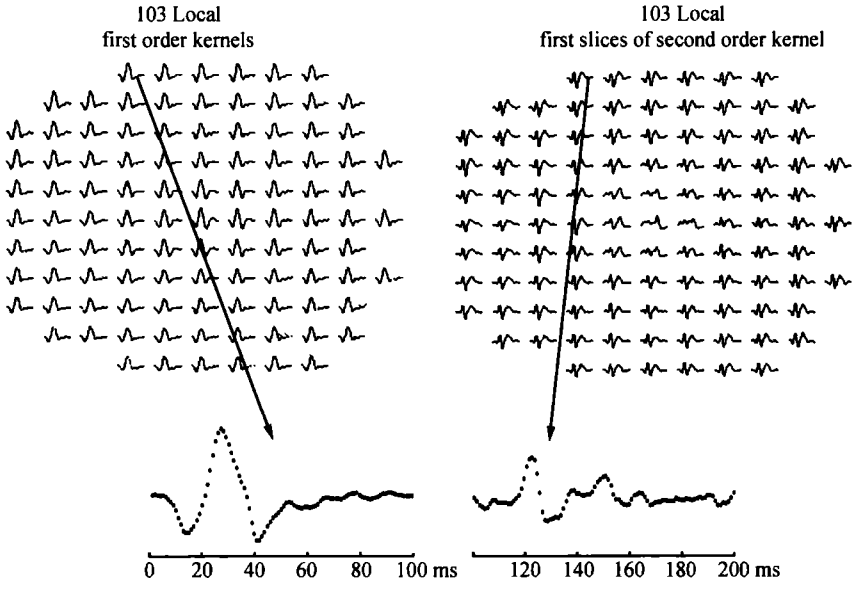

Fig. 1. Waveform of each local mfERG response covered 200 ms long. This figure is to illustrate the method in data acquisition. Waveforms did not come from experimental data. Description of the image is given in the text.

where i refers to the sequence number of the local response, from 1 to 103. The feature subset $\boldsymbol{X}_i$ is d-dimensional, which means that $\boldsymbol{X}_i$ is composed of d features selected from the 200 features contained in the local response array. The $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ are the between-class scatter matrix and the within-class scatter matrix of $\boldsymbol{X}_i$, respectively. According to the definition of the criterion function, feature subset with bigger $J(\boldsymbol{X})$ value should have better classification performance.

To evade computational complexity in full searching of the best d-dimensional $\boldsymbol{X}_i$ within the 200 original features, the sequential forward selection (SFS) algorithm[10] was used to find the close-optimal feature subset $\boldsymbol{X}_i^*$ to maximize $J(\boldsymbol{X}_i^*)$.

2.2 Training of linear classifier

After all feature subsets $\boldsymbol{X}_i$ were selected from the 1st to 103rd local mfERG responses, we used the two-category linear classifier to classify the experimental subjects. Discriminant function of the classifier for each selected feature vector $\boldsymbol{X}_i$ is $g_i(\boldsymbol{X}_i)$

$$\begin{aligned} &g_i(\boldsymbol{X}_i) = \boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{X}_i + w_0^i \\ &\boldsymbol{W}_i = [w_{i1}, w_{i2}, \cdots, w_{id}]^{\mathrm{T}} \\ &i = 1, 2, \cdots, 103 \end{aligned} \tag{2}$$

where $\boldsymbol{W}_i$ is the weight vector, and w_0^i is the threshold weight. The decision rule of the linear classifier is

$$\begin{cases} \text{if } g_i(\boldsymbol{X}_i) \geqslant 0, & \boldsymbol{X}_i \in C1 \\ \text{if } g_i(\boldsymbol{X}_i) < 0, & \boldsymbol{X}_i \in C2 \end{cases} \tag{3}$$

where $C1$ and $C2$ are labels of the two classes to be classified.

The principles of Fisher's linear classifier[11,12] was used to compute parameters $\boldsymbol{W}_i$ and w_0^i of the discriminant function $g_i(\boldsymbol{X}_i)$. To obtain these parameters, Eqs. (4)—(6) were applied to the $C1$ and $C2$ classes of experimental subjects.

$$\boldsymbol{W}_i = \boldsymbol{S}_w^{-1}(m_1 - m_2) \tag{4}$$

where $\boldsymbol{S}_w$ is the within-class scatter matrix of $\boldsymbol{X}_i$, m_1 and m_2 are sample mean of the $C1$ and $C2$ class, respectively. Giving $\boldsymbol{W}_i$, the threshold weight w_0^i is

$$w_0^i = -\frac{(n_1 m_1' + n_2 m_2')}{(n_1 + n_2)} \tag{5}$$

where

$$m_j' = \frac{1}{n_j}\sum_{\boldsymbol{X}_i \in C_j} \boldsymbol{W}_i^{*\mathrm{T}}\boldsymbol{X}_i \quad j = 1, 2 \tag{6}$$

and n_1 and n_2 are samples in classes $C1$ and $C2$, respectively.

2.3 Classification results for all local responses

Given mfERG results of $C1$ and $C2$ classes of experimental subjects, there should be 103 selected d-dimensional feature vectors $\boldsymbol{X}_i$ ($i = 1, 2, \cdots, 103$) and discriminant functions $g_i(\boldsymbol{X}_i)$, ($i = 1, 2, \cdots, 103$), each corresponding to a local mfERG response. This means that for one subject to be classified, there are 103 classification results when every function $g_i(\boldsymbol{X}_i)$ is applied to the local mfERG response corre-

spondently. To obtain the final classification decision of the subject, a vote value EV was computed to combining the 103 classification outputs. The vote rule is

$$EV = \sum_{i=1}^{103} p_i$$

where,
$$\begin{cases} p_i = 1 & \boldsymbol{X}_i \in C1 \\ p_i = -1 & \boldsymbol{X}_i \in C2 \end{cases} \quad (7)$$

The classification rule of the subject is: if $EV \geqslant 0$, the subject is classified to $C1$ class, else if $EV < 0$, then the subject is classified to class $C2$.

2.4 Classification error estimation

Applying the classification method to each test sample, the classification error rate could be estimated by

$$\text{Error rate} = \frac{\text{Misclassified test samples}}{\text{All test samples}} \quad (8)$$

3 Result

The method of mfERG data analysis discussed before was used on those four classes of subjects (Tables 1 and 2). Those subjects' retinal state changed from healthy (Normal) to proliferative retinopathy (PDR), according to progression of the diabetic retinopathy. To classify the experimental subjects, Normal, NDR, NPDR and PDR, five two-category classification problems were solved (No. 1—5 in Table 2).

Also shown in Table 2, about a half experimental subjects were randomly selected to form the training data set for the feature selection and classifier training steps, and the rest of them were grouped as a test data set to compute error rates of the classification method.

Firstly, the feature selection method was applied to the training set of the four classes of subjects. The dimension of the target feature subset was set to $d = 6$. And 103 feature subsets ($\boldsymbol{X}_i$, $i = 1, 2, \cdots, 103$) were selected from the feature space, each corresponding to a local mfERG response.

Having the feature subset selected, for every classification problem in Table 2, coefficients of the 103 discriminant function $g_i(\boldsymbol{X}_i)$ were trained following formulas (4)—(6), over the data in training set.

To test performance of the trained classifiers, the function $g_i(\boldsymbol{X}_i)$ were applied to each subject in the test set. The EV value was given by (7) to combine the outputs of all $g_i(\boldsymbol{X}_i)$, and the subject was classified according to sign of the EV value. The classification error rate was estimated by (8). The average EV values among all subjects in each test set are also given in Table 2. Table 3 gives the confusion matrix when classifying the test subjects for all the classification problems in Table 2.

Table 2. Classification results for the four categories of experimental subjects

No.	Classification target ($C1/C2$)	Training set		Test set		Classification results	
						EV (average value)	Error rate (%)
1	Normal	15	44	16	45	27	2.2
	NDR	29		29		−35	
2	Normal	15	37	16	39	65	0
	NPDR	22		23		−68	
3	NDR	29	51	29	52	59	0
	NPDR	22		23		−53	
4	NDR	29	42	29	42	75	0
	PDR	13		13		−81	
5	NPDR	22	35	23	36	35	8.3
	PDR	13		13		−18	

Table 3. Confusion matrices for the classification problem in Table 2

No.	True labels	Classification		Total
		Normal	NDR	
1	Normal	16	0	16
	NDR	1	28	29
		Normal	NPDR	
2	Normal	16	0	16
	NPDR	0	23	23
		NDR	NPDR	
3	NDR	29	0	29
	NPDR	0	23	23
		NDR	PDR	
4	NDR	29	0	29
	PDR	0	13	13
		NPDR	PDR	
5	NPDR	21	2	23
	PDR	1	12	13

It can be seen that No. 2—No. 4 experiments got zero classification error rate. And the error rate in discriminating the Normal subjects from the NDR subjects was 1/45 (2.2%). Correspondingly shown in Table 3 that one Normal subject was mistakenly classified to NDR class. In classification between the NPDR and PDR subjects (No. 5 experiment in Table 2), the error rate was 3/36 (8.3%). The misclassified subjects were two NPDR subjects being regarded as PDR, and one PDR subject being regarded as NPDR.

There is another noticeable result in Tables 2 and 3. The NDR eyes were discriminated from the Normal eyes at 2.2% error rate. As we knew, the NDR subjects discussed here referred to those diabetic eyes having no ophthalmoscopically detectable retinopathy, which means that fundus photograph of all those NDR subjects did not show any apparent abnormities caused by diabetes.

Good discrimination of these two kinds of experimental subjects confirmed the existence of retinal dysfunction in diabetic eyes without observable retinopathy symptoms. And these retinal dysfunctions could be identified by analyzing the mfERG data at a very high accuracy rate.

According to the classification results, our data analysis method provided a great classification performance for the four groups of experimental subjects. This method included defining the amplitude value of each point on the mfERG responses as a potential feature, selecting the most discriminative feature subset from all features for each local mfERG response, and training the linear classifiers. For one test subject, the trained linear classifiers were applied on all of its local mfERG responses. Therefore the subject could be classified by voting on all the local classification results.

4 Discussion

Besides high accuracy, a good disease classification method should have good stability and ability to predict the progression of the disease. Two tests are presented in this section to discuss the properties of our DR classification method.

4.1 Stability of the method

Firstly, because the training and test data set in Table 2 were formed by choosing samples from the original data set randomly, and intersection between them was kept empty, the low classification error rates preliminary proved stability of the method. To further testify this judgment, classifications between the NDR and NPDR samples were performed when different groups of subjects were chosen to form the training and test data set.

According to mfERG data obtained from 58 NDR and 45 NPDR subjects, shown in Table 1, 29 NDR and 22 NPDR subjects were randomly selected to compose the training set, and for each of the 103 local mfERG responses the best 6-dimensional feature vector and the linear classifier were computed based on the training set. After that, the test set, including 29 NDR and 23 NPDR subjects, was randomly selected from the 58 NDR and 45 NPDR subjects for 10 times. The classification result for each of the 10 test set is shown in Table 4, so are the average and standard deviation values.

Table 4. Classification between NDR and NPDR under different test set grouping

No.	*EV* (average value)		Error rate (%)
	NDR	NPDR	
1	67	−46	0
2	60	−32	0
3	64	−37	0
4	37	−25	7.7
5	63	−53	0
6	66	−35	0
7	62	−32	0
8	43	−27	5.8
9	47	−30	1.9
10	40	−35	3.8
Average value ± SD	54.9 ± 11.7	−35.2 ± 8.5	1.92 ± 2.7

According to Table 4, the classification error rates kept low in 10 tests with different data grouping. This confirmed our judgment for good stability of our DR classification method. In addition, comparing the EV value and corresponding error rate in Table 4, we found that the closer the EV value approached to zero, the higher probability of classification error happened (No.4 and 8 in Table 4). This relationship suggested the evaluation ability of the EV value on risk of error classification.

4.2 Relationship between the classifiers and the progression of retinopathy

To test whether our method could indicate the progression of DR, No. 4 classification problem in Table 2, discriminating between the NDR and NPDR samples, was re-designed. Some extra samples which did not belong to the NDR or NPDR classes were added to the test set of that classification problem. And the classifiers trained by the training set of NDR and NPDR data were applied to the new test set.

Ten mfERG records were randomly selected from the Normal class and were added to the test set, grouped by NDR and NPDR individuals, in Table 3 (No.3). The confusion matrix in classifying the new test set is shown in Table 5 (No.1). All ten extra Normal subjects were recognized as NDR subjects. Ten PDR individuals were appended to the test set Table 3 (No.3), and the classification result for this test set is shown in Table 5 (No.2).

Table 5. Confusion matrix of the classification between NDR and NPDR when new testing samples were added

No.	True labels	Classification		Total
		NDR	NPDR	
1	NDR	29	0	29
	NPDR	0	23	23
	Normal	10	0	10
		NDR	NPDR	
2	NDR	29	0	29
	NPDR	0	23	23
	PDR	0	10	10

Four classes of the experimental subjects discussed in this paper had different retinal dysfunctional level from none (Normal) to severe (PDR). The classification result in Table 5 shows that the DR classification method proposed in this paper is consistent with progression of the retinal dysfunction caused by diabetic retinopathy.

References

1 Brailsford SC, Davies R, Canning C, et al. Evaluating screening policies for the early detection of retinopathy in patients with non-insulin dependent diabetes. Health Care Management Science, 1998, 1: 115—124

2 Sutter EE and Tran D. The field topography of ERG components in man. I. The photopic luminance response. Vision Res, 1992, 32: 433—466

3 Hood DC. Assessing retinal function with the multifocal technique. Prog Retin Eye Res, 2000, 19: 607—646

4 Sutter EE. The fast m-transform: a fast computation of cross-correlations with binary m-sequences. Soc Ind Appl Math, 1991, 20: 686—694

5 Palmowski AM, Sutter EE, Bearse MA, et al. Mapping of retinal function in diabetic retinopathy using the multifocal electroretinogram. Invest Ophthalmol Vis Sci, 1997, 38: 2586—2596

6 Fortune B, Schneck ME and Adams AJ. Multifocal electroretinogram delays reveal local retinal dysfunction in early diabetic retinopathy. Invest Ophthalmol Vis Sci, 1999, 40: 2638—2651

7 Shimada Y, Li Y, Bearse MA, et al. Assessment of early retinal changes in diabetes using a new multifocal ERG protocol. Br J Ophthalmol, 2001, 85: 414—419

8 Bearse MA, Han Y, Schneck ME, et al. Retinal function in normal and diabetic eyes mapped with the slow flash multifocal electroretinogram. Invest Ophthalmol Vis Sci, 2004, 45: 296—304

9 Early Treatment Diabetic Retinopathy Study Research Group. Early photocoagulation for diabetic retinopathy: ETDRS report No. 9. Ophthalmology, 1991, 98: 766—785

10 Liu H and Motoda H. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic, 1998

11 Webb A. Statistical Pattern Recognition. London: Arnold, 1999

12 Duda R, Hart P and Stork D. Pattern Classification. New York: Wiley, 2000