IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 5, 2017, accepted February 17, 2017, date of publication February 20, 2017, date of current version March 15, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2671918

Automatic Microaneurysm Detection Using the Sparse Principal Component Analysis-Based Unsupervised Classification Method

WEI ZHOU[1,2], CHENGDONG WU[1,2], DALI CHEN[1], YUGEN YI[3], AND WENYOU DU[1]

[1]College of Information Science and Engineering, Northeastern University, Shenyang 110004, China
[2]Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110004, China
[3]School of Software, Jiangxi Normal University, Nanchang 330022, China

Corresponding authors: W. Zhou (zhouweineu@outlook.com) and Y. Yi (yiyg510@jxnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61273078, Grant 61471110, and Grant 61602221, in part by the Foundation of Liaoning Educational Department under Grant L2014090, and in part by the Fundamental Research Funds for the Central Universities under Grant N140403005.

ABSTRACT Since microaneurysms (MAs) can be seen as the earliest lesions in diabetic retinopathy, its detection plays a critical role in the diabetic retinopathy diagnosis. In recent years, many machine-learning methods have been developed for MA detection. Generally, MA candidates are first identified and then a set of features for these candidates are extracted. Finally, machine-learning methods are applied for candidate classification. In this paper, we present a novel unsupervised classification method based on sparse posterior cerebral artery (PCA) for MA detection. Since it does not have to consider a non-MA training set, the class imbalance problem can be avoided. Furthermore, effective features can be selected due to the characteristic of sparse PCA, which combines the elastic net penalty with the PCA. Meanwhile, a single T^2 statistic is introduced, and the control limit can be determined for distinguishing true MAs from spurious candidates automatically. Experiment results on the retinopathy online challenge competition database show the effectiveness of our proposed method.

INDEX TERMS Diabetic retinopathy, microaneurysm detection, sparse PCA, unsupervised classification.

I. INTRODUCTION

Diabetic retinopathy (DR), also known as diabetic eye disease [1] and has become a major cause of blindness among the middle-aged population [2]. Early diagnosis through regular screening is recommended to diabetic patients, which can help them prevent blindness and visual loss. However, a large amount of diabetic patients need to be screened annually, which poses a heavy workload for ophthalmologists. Therefore, developing an automatic DR screening system is necessary, which can not only reduce the workloads of ophthalmologists, but also improve the accuracy of detection [3].

Color fundus images with low-cost and patient friendliness are widely used for automatic DR detection [4]. In the color fundus images, the signs of DR contain red lesions such as MAs and hemorrhages, bright lesions such as hard exudates and cotton wool spots. Among the lesions, MAs, which appear as small circular dark spots on the surface of the retina (see Figure 1), can be regarded as the earliest visible lesion in diabetic retinopathy [5]. Therefore, MA detection plays a critical role in DR detection and we focus on it within this paper.

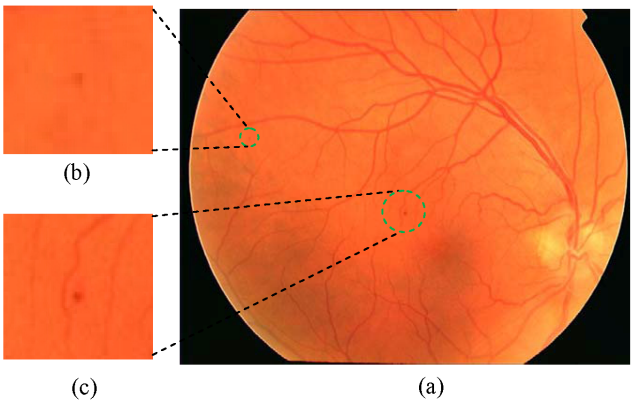


FIGURE 1. A digital fundus image with two marked MAs (a) a RGB fundus image; (b) the size of MA radius with five pixels; (c) the size of MA radius with ten pixels.

Numerous approaches have been proposed for MA detection, they usually start with image preprocessing to remove the uneven illumination and reduce noises in the retinal images. Then, all the possible MA candidates are located by an initial detection method. Finally, MA classification based on features computed on each candidate [6].

The earliest MA detection method was proposed by Baudoin *et al.* [7]. They used mathematical morphology approach to detect the microaneurysms in fluorescein angiogram. In this approach, MA candidates can be determined by applying 12 morphology top-hat transformations with linear structuring elements at different orientations to green channel of the image. After that, two variants of morphological top-hat transformation methods for extracting MAs within fluorescein angiograms were developed by Spencer *et al.* [8] and Frame *et al.* [9]. Although using the fluorescein angiograms can improve the contrast between the fundus and their background, the usage of intravenous contrast agents is not applicable for everyone, such as the pregnant woman [10]. Hence, it cannot be widely used in public DR screening programs. Besides, mathematical morphology based approaches mainly depend on the choosing of structuring elements, when changing their size and shape, it may increase false positives or decrease true positives.

Apart from the above mentioned MA detection approaches, a number of template matching based algorithms have been proposed for MA detection. In [11], a local template matching in the wavelet domain has been used for detecting MAs. The problem of illumination variations or high-frequency noise can be avoided effectively in this approach. In addition, Zhang *et al.* [12] employed Multi-scale Gaussian Correlation Coefficients (MSCF) method to detect MAs. In this work, MA candidates can be detected by computing the maximum correlation coefficient with five different Gaussian kernels for each pixel. And then 31 features were extracted for each candidate. Finally, true MAs can be identified by specifying the thresholds for each feature directly. However, setting threshold for all the features depends on prior knowledge of experts [12]. Also, simplify taking all features into account is not appropriate. It is inevitably to introduce some irrelevant or redundant features, which not only deteriorates classification performance, but also is time-consuming. So how to choose the useful subset of features for candidate classification should be considered.

Several machine learning algorithms are used for MA detection. Niemeijer *et al.* [13] presented a hybrid scheme which combined morphological top-hat transform with a Knn classifier for MA detection. After that Sánchez [14] proposed an approach which integrated the Gaussian mixture model with a logistic regression classification into a unified framework for MA detection. Furthermore, Zhang *et al.* [15] developed a MA detection method, which combined the Dictionary Learning (DL) with Sparse Representation Classification (SRC). In their method, firstly, Multi-scale Gaussian Correlation Coefficients filtering was applied to locate all the possible candidates. And then two dictionaries were learned based on the MA candidates and non-MA candidates. Finally, true MAs were classified by SRC.

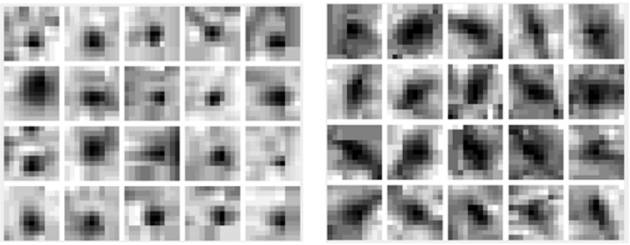


FIGURE 2. Examples of MA and non-MA. (a) MA samples; (b) non-MA samples.

A common problem occurs in the aforementioned methods that features for non-MAs vary in a wide range. In order to collect enough training set representing all non-MAs, the number of non-MA samples has to be very large. An example is shown in Figure 2, which illustrates a series of labeled image patches (MAs and non-MAs) extracted by [15]. From Figure 2, we can see that the MAs (Figure 2(a)) are similar to each other, appeared as circle structures, but the non-MAs (Figure 2(b)) are different from one another without a unified structure. In this scenario, how to collect non-MA training set is quite subject, large training set is not only time-consuming but also will cause the class imbalance problem [10].

The rest of the paper is organized as below: in Section II, we describe several preliminary works for the proposed method. It consists of the following three phases: preprocessing, candidate extraction and feature extraction. The sparse PCA based classification method for MA detection is proposed in Section III. Experiments are carried out on ROC [16] database and results are reported in Section IV and Section V concludes the paper.

II. PRELIMINARIES

As stated in [17], the MA appears more contrast in green channel of fundus images. Firstly, a preprocessing is applied to the green channel fundus image for reducing noise and improving contrast. Secondly, multi-scale Gaussian correlation coefficients method is used to extract all the possible MA candidates. Lastly, a total of 34 features based on shape, brightness and Gaussian filtering are extracted for each candidate forming the feature matrix.

A. PREPROCESSING

The large luminosity, poor contrast and noise always occur in retinal fundus images, which affect seriously the diagnostic process of DR and the automatic lesions detection, especially for MA [11]. In order to address these problems and make a suitable image for MA detection, firstly, extracting the green channel of original image, in which the MAs have the higher contrast with their background [17]. After that, contrast limited adaptive histogram equalization (CLAHE) [18] method is applied to the green channel image for making the

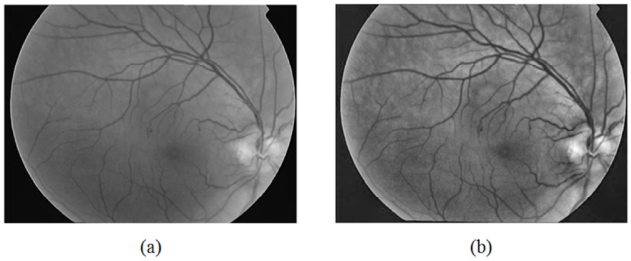


FIGURE 3. Preprocessing. (a) the green band of original image; (b) image result after preprocessing.

hidden features more visible. At the end, Gaussian smoothing filter with a width of 5 and a standard deviation of 1 is also incorporated to enhanced image for reducing the effect of noise further. An instance for describing an original green channel retinal image and the final result image processed by the above steps are shown in Figure 3.

B. CANDIDATE EXTRACTION

Multi-scale Gaussian Correlation Coefficients (MSCF) proposed by Zhang *et al.* [12] is applied to extract MA candidates. In their method, MA candidates are detected by computing the correlation coefficient between Gaussian function with five different Gaussian kernels and the distributions of its grayscale image. The details of this method are listed as below:

Firstly, a non-linear filter with five different Gaussian kernels is used for calculating a correlation coefficient of each pixel. The maximum coefficient at each pixel location among the five responses is selected to form the final response. Since MAs vary in size, the final response with different Gaussian kernels allows detecting MAs of varying sizes (see Figure 4 (b)). Here, we denote Gaussian function and the gray distribution of MAs by the variables of A and B respectively. The correlation coefficient can be defined as:

$$coeff_{AB} = \frac{\sum_m \sum_n \left(A_{mn} - \bar{A}\right)\left(B_{mn} - \bar{B}\right)}{\sqrt{\left(\sum_m \sum_n \left(A_{mn} - \bar{A}\right)^2\right)\left(\sum_m \sum_n \left(B_{mn} - \bar{B}\right)^2\right)}} \tag{1}$$

where $\bar{A}$ and $\bar{B}$ are the mean values of A and B, and the values of correlation coefficient range from 0 to 1.

Secondly, in order to reduce the number of microaneurysm candidates in final response. A threshold T which ranges from 0.1 to 0.9 with an interval of 0.1 is applied to eliminate the candidates with low coefficients. Since the MAs do not appear on the vasculature, any candidates on the vasculature need to be removed [19] (the extracted vasculature map is shown in Figure 4 (c)). In addition, the size and shape of detected MA candidates do not represent the true MAs. Hence, region growing [8], [13] is used for solving this issue. In the region growing, the background image I_{bg} can be obtained by applying mean filter to green channel image I_{green}. An adaptive threshold t based on the dynamics

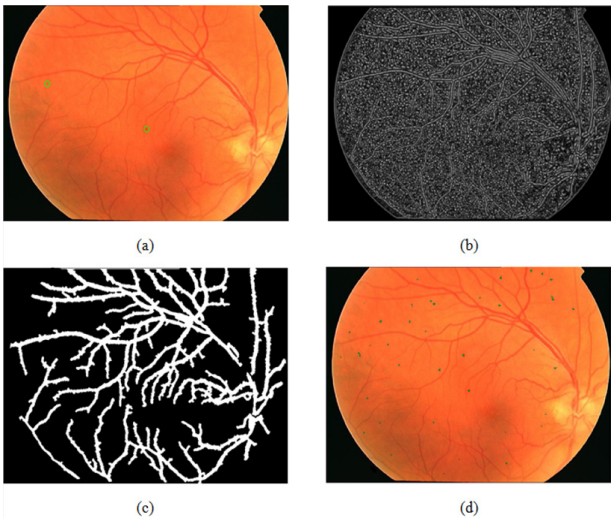


FIGURE 4. Candidate extraction. (a) the retinal image with annotated mircroaneurysms; (b) the final response of multi-scale correlation filtering; (c) the blood vessel map of (a); (d) the final result map of region growing.

is given by Eq. (2):

$$t = I_{darkest} - \beta \cdot \left(I_{darkest} - i_{bg}\right) \tag{2}$$

where $I_{darkest}$ denotes the lowest intensity for each candidate region in the I_{green}, i_{bg} is its background intensity at the same location, β is a constant value ranging from 0 to 1, which is set to 0.5 here.

Region growing starts from the point of $I_{darkest}$ in each candidate region and continues until no more connected pixels are higher than threshold. Considering the size of MA is less than 120 pixels [12], if the area of every resultant connected component is larger than 120 pixels, it will be discarded.

Finally, the remaining resultant connected components can be regarded as the final MA candidate regions (see Figure 4 (d)). A systematic overview of candidate extraction is shown in Figure 4.

C. FEATURE EXTRACTION

Considering the fact that MAs appear as circle structure, red dense regions with low intensity in fundus images, therefore, we extract the following features for each candidate region with the aim of distinguishing the MA candidates from the non-MA candidates further. Here, a total of 34 different features are extracted for each candidate. Those features can be divided into three categories: shape-based, intensity and color based and Gaussian filtering based.

Shape-Based Features

(1-2) Area and Perimeter: $a = \sum_{j \in \Omega} 1$ and p, where Ω is the set of pixels in the candidate region.

(3-5) Axis lengths: l and w are major axis lengths and minor axis lengths of the candidate, and aspect ratio: $r = {}^{l}/_{w}$.

(6) Circularity: $c = {}^{p^2}/_{(4 * \pi * a)}$.

(7) Compactness: $v = \sqrt{{}^{\sum_{j=1}^{n} \left(d_j - \bar{d}\right)}/_{n}}$ where d_j is the distance from the centroid of the object to its j-th boundary

pixel and $\bar{d}$ is the mean of all the distances from the centroid to all the edge pixels. Here n is the number of edge pixels.

Intensity and Color Based Features

(8-9) Total intensity: candidate region in I_{green}, $i_{green} = \sum_{j\in\Omega} g_j$ and candidate region in I_{sc}, $i_{sc} = \sum_{j\in\Omega} s_j$.

(10-11) Mean intensity: candidate region in I_{green}, $m_{green} = {i_{green}}/{a}$ and candidate region in I_{sc}, $m_{sc} = {i_{sc}}/{a}$.

(12) The mean contrast of edge pixels: $C = \frac{\sum_{j\in A} g_j}{num_A} - \frac{\sum_{j\in B} g_j}{num_B}$ denotes a pixel contrast, where A is the 8 neighboring pixels of current pixel with high intensity and B is the remaining 8 neighboring pixels of the A, num_A is the number of pixels in A.

(13) The standard deviation of edge pixels: $\sigma_C = \frac{\sum_{j\in Z}\sqrt{(C_j - C)^2}}{num_Z}$, where Z is the set of edge pixels and num_Z is the number of pixels in Z.

(14-15) Normalized intensity: candidate region in I_{green}, $NI_{green} = ({1}/{\sigma})\left(i_{green} - \bar{x}\right)$ and candidate region in I_{sc}, $NI_{sc} = ({1}/{\sigma})(i_{sc} - \bar{x})$ where σ and $\bar{x}$ are the standard deviation and average pixel value of I_{bg} and I_{sc} respectively.

(16-17) Normalized mean intensity: candidate region in I_{green}, $NM_{green} = ({1}/{\sigma})\left(m_{green} - \bar{x}\right)$ and candidate region in I_{sc}, $NM_{sc} = ({1}/{\sigma})(m_{sc} - \bar{x})$.

(18-19) The mean intensity and standard deviation of background: $M_{bg} = \frac{i_{bg}}{num_{bg}}$ and σ_{bg}, where i_{bg} is a set of background pixels of each candidate region.

(20-21) The mean intensity and standard deviation in CLElab color space.

(22-23) The minimum intensity and maximum intensity in CLElab color space.

Gaussian Filtering Based Features:

(24-31) Mean and standard deviation: Gaussian filtering responses of I_{green} with different scales $\sigma = 1, 2, 4, 8$.

(32-34) Correlation coefficient of candidates: maximum, minimum and average.

Some features listed above require I_{sc}. Next, we will describe how to calculate it. I_{sc} is I_{green} with the background I_{bg} removed calculated as $I_{green} - I_{bg}$. In summary, a total 34 features (7 shape features + 16 color and intensity features +11 Gaussian filtering features) have been extracted for each candidate region.

III. SPARSE PCA BASED CLASSIFICATION FOR MA DETECTION

The proposed sparse PCA based classification method is depicted in this section. Firstly, we will make a short review of PCA and sparse PCA. Next, T^2 statistic of MA and its control limit are given. Finally, a synthetic example is utilized for validating the accuracy of classification.

A. PRINCIPAL COMPONENT ANALYSIS

PCA is a linear dimensionality reduction method. Its basic principle is to maximum the variance of projections on new directions. It can be determined by a group of orthogonal vectors called the load vectors and the sequence of vectors is determined by variance of projections on load vectors. Given a set of training samples $X \in R^{N\times m}$ with N samples and m variables, the objective function of optimization problem can be written as follows [20].

$$\max_{v\neq 0} \frac{v^T X^T X v}{v^T v} \tag{3}$$

where $v \in R^m$. The steady solution of Eq. (3) can be calculated by Singular Value Decomposition (SVD).

$$\frac{1}{\sqrt{n-1}} X = U \Sigma V^T \tag{4}$$

where $U \in R^{n\times n}$ and $V \in R^{m\times m}$ are all unitary matrices, $\Sigma \in R^{n\times m}$ consists of non-negative singular value in descending order ($\sigma_1 \geq \sigma_2 \cdots \geq \sigma_{\min(m,n)} \geq 0$) along the main diagonal, other elements in Σ are zeros. Orthogonal column vectors in V correspond to load vectors. Projection of X on i-th column has the variance equals to σ_i^2. Select the first a columns in V to compose a new load matrix $P \in R^{m\times a}$ which is also called the principal component subspace (PCS).

$$T = XP \tag{5}$$

A new sample vector x can be projected on the PCS:

$$t = P^T x \in PCS \tag{6}$$

About the value a, it is given by Cumulative Percent Variance (CPV) which is calculated as follows:

$$CPV(a) = 100 \left[\frac{\sum_{j=1}^{a} \lambda_j}{\sum_{j=1}^{m} \lambda_j} \right] \% \tag{7}$$

The CPV is the percent variance captured by the first a principle components (PCs); m is the total number of eigenvalues. It is very subjective what value for a is suitable, Sergio Valle et al., discussed this problem in their article [21].

B. SPARSE PCA

There are several variables (or features) linear relevant with each other, but, some of them are observed with relatively large noises. We call this kind of variables as weak relevant. Since the main purpose of PCA is to obtain the maximum variance on certain loading vectors, some principal components will be inevitably represented by such kind of weak relevant variables which will decrease the detection precision.

LASSO-based PCA performs the maximization of objective function (3) under the extra L1 norm constraint:

$$\sum_{j=1}^{m} |v_j| \leq sparsity \tag{8}$$

Here v_j represent the jth element of the load vector $j \in \{1, 2, \ldots, m\}$, *sparsity* represents the number of the non-zero elements in a load vector. With this constraint, principal components can be represented with less original variables. Several algorithms have been developed for solving this

problem. Among them, one of the most effective and representative methods is proposed by [22]. Basically, PCA can first be recast exactly in terms of a (ridge) regression problem. And then, the L1 norm constraint can be introduced, changing this regression problem to an elastic-net regression. A short review will be given as follows.

Consider the fact that each PC is a linear combination of the m variables, thus its sparse loadings can be obtained by regressing the PC on the m variables. For each i, let $Z_i = XV_i$ denote the ith principal component and λ be a positive constant. The ridge estimates $\hat{\mu}_{ridge}$ is given by:

$$\hat{\mu}_{ridge} = \arg\min_{\mu} \|Z_i - X\mu\|^2 + \lambda \|\mu\|^2 \tag{9}$$

where $\hat{\mu}_{ridge}$ can be obtained by $\hat{v} = \frac{\hat{\mu}_{ridge}}{\|\hat{\mu}_{ridge}\|^2}$, and $\hat{v} = V_i$ is the ith loading.

Now, let's discuss the contribution of L2 norm in the objective function and the selection of λ. When there are many correlated variables in a linear regression model, their coefficients can be poorly determined and exhibit high variance. A widely large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By imposing a size constraint on the coefficients, as the L2 norm in ridge regression, this problem is relieved with a relatively small value of λ. Besides, when $m > N$, if $\lambda = 0$, ordinary multiple regression has no unique solution that is exactly V_i, it is the same when $N > m$ and X is not a full rank matrix. With the L2 norm, we can always give the unique solution in all situations.

Next, we will add the L1 penalty to Eq. (9) and obtain the following objective function:

$$\hat{\mu} = \arg\min_{\mu} \|T_i - X\mu\|^2 + \lambda \|\mu\|^2 + \tau \|\mu\|_1 \tag{10}$$

Here, $\|\mu\|_1 = \sum_{j=1}^{m} |\mu_j|$ is the L1 norm of μ. And $\hat{V}_i = \frac{\mu}{\|\mu\|}$ is an approximation of the loading vectors. τ is a tradeoff parameter controls the sparsity of loading vectors, bigger τ corresponds to less non-zero elements in loading vectors, In practice, we can choose a proper τ by specifying the sparsity which will be discussed later. Solving this objective function, we can achieve the sparse PCA and obtain the sparse loading vectors $\hat{V}_i$. As shown in Eq. (10), it is a penalized least squares problem using the elastic net penalty, which can be solved with an efficient LARS-EN algorithm [23]. Details of the elastic net penalty and LARS-EN algorithm can be found in the reference.

Sparse PCA has several advantages, the motivation introducing sparse PCA for MA detection are summarized as follows:

1) L1 norm constraint is introduced to achieve the variance-sparsity trade-off and weak relevant variables in feature matrix can be selected out by sparsity. Besides, sparse loadings also make the principal components easy to interpret by a human.
2) L2 norm constraint is introduced to give the unique solution in all situations. Meanwhile, multicollinearity can also be alleviated with this constraint.

In our work, we mainly focus on the detection performance using sparse PCA. So, we will provide a synthetic example to evaluate the effectiveness compared with ordinary PCA. And PR curve is used to verify the classification performance. Next, we will introduce the Index for detection.

C. INDEX AND CONTROL LIMIT

There are several statistics applied to multivariable statistical process monitoring. Typically Hotelling's T^2 statistic is used to represent the variability in the PCS [24]. It is defined as follows:

$$T^2 = x^T P \Lambda^{-1} P^T x \tag{11}$$

Here, $\Lambda = \Sigma^T \Sigma$. It is also assumed that the sample vector x follows a multivariate normal distribution. We have to validate the assumption before using our method. We perform the Single sample Kolmogorov-Smirnov goodness-of-fit hypothesis test. The null hypothesis is that each feature has a standard normal distribution. The alternative hypothesis is that each feature does not have that distribution. In our paper, the significance level is set at 0.05 and the decision to reject the null hypothesis is based on comparing the p-values with the significance level. The p-values of Single sample Kolmogorov-Smirnov hypothesis test for all extracted 34 features are listed in Table 1. As can be seen, all p-values are bigger than 0.05 which means that all features basically follow the normal distribution.

TABLE 1. The *p*-values of single sample Kolmogorov-Smirnov hypothesis test with 0.05 significance level.

Feature	p-values	Feature	p-values	Feature	p-values
1	0.85	13	0.58	25	0.30
2	0.74	14	0.22	26	0.16
3	0.21	15	0.34	27	0.21
4	0.27	16	0.95	28	0.22
5	0.58	17	0.21	29	0.65
6	0.45	18	0.31	30	0.28
7	0.54	19	0.36	31	0.12
8	0.49	20	0.83	32	0.91
9	0.55	21	0.64	33	0. 12
10	0.58	22	0.82	34	0.52
11	0.21	23	0.41		
12	0.68	24	0.72		

The effect of T^2 can be illustrated in Figure 5 which reveals that original input variables are linear correlation. Figure 5(b) illustrates that after PCA decomposition score vectors on principal components have the largest variance. Figure 5(c) illustrates that new score vectors follow the multivariate normal distribution with the same mean values equal to 0 and the same variances equal to 1. Under the condition that the data are normal and follow a multivariate normal distribution,

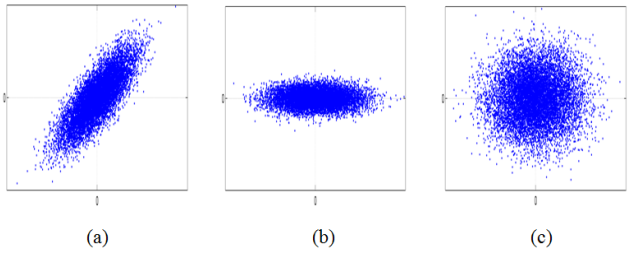


FIGURE 5. Data transformation process. (a) data on original space; (b) the first two SPCs; (c) score divided by corresponding eigenvalues.

the T^2 statistic is related to an F distribution [25]:

$$\frac{N\,(N-a)}{a\left(N^2-1\right)}T^2 \sim F_{a,N-a} \tag{12}$$

where $F_{a,N-a}$ is an F distribution with a and $N-a$ degrees of freedom. a is the number of retained principal components, and N is the number of samples. For a given significance level α, the new candidate is considered as true MA if

$$T^2 \leq T_\alpha^2 \equiv \frac{a(N^2-1)}{N(N-a)}F_{a,N-a;\alpha} \tag{13}$$

In the following part, we will validate the effectiveness of sparse PCA based detection method, the way choosing the proper sparsity will be also discussed.

D. A SYNTHETIC EXAMPLE VALIDATION

In this part we will validate the effectiveness of sparse PCA for classification compared with PCA. The PR curve will be introduced to evaluate the classification performance and determine the sparsity. Now, let us introduce the linear model of our synthetic example:

$$\begin{cases} x_1 = rand_1 \\ x_2 = rand_2 \\ x_3 = 1.3 \times x_1 + 0.2 \times rand_3 \\ x_4 = 0.8 \times x_1 + 0.2 \times rand_4 \\ x_5 = 2 \times x_2 + 0.2 \times rand_5 \\ x_6 = x_2 + 0.2 \times rand_6 \\ x_7 = 0.2 \times x_2 + rand_7 \end{cases}$$

$$\begin{cases} x_8 = 0.2 \times x_1 + rand_8 \\ x_9 = 0.2 \times x_2 + 5 \times rand_9 \\ x_{10} = 0.8 \times x_1 + 5 \times rand_{11} \\ x_{11} = 0.1x_1 + 0.7x_2 + 5 \times rand_{13} \\ x_{12} = 0.1x_1 + 0.7x_2 + 5 \times rand_{13} \\ x_{13} = 5 \times rand_{13} \\ x_{14} = 5 \times rand_{13} \end{cases} \tag{14}$$

$$\begin{cases} x_1 = 6 \times rand_{15} \\ x_2 = 6 \times rand_{16} \\ x_{10} = x_{10} + 3 \\ x_{11} = x_{11} - 3 \\ x_{12} = x_{12} + 1 \\ x_{13} = x_{13} - 1 \\ x_{14} = x_{14} - 5 \end{cases} \tag{15}$$

TABLE 2. Loadings of the First Three PCs by Ordinary PCA and SPCs by Sparse PCA

	PC1	PC2	PC3	SPC1	SPC2	SPC3
x_1	0.47	0.30	-0.02	-0.57	0	0
x_2	-0.29	0.46	-0.10	0	0.57	0
x_3	0.47	0.30	-0.02	-0.57	0	0
x_4	0.47	0.29	-0.03	-0.57	0	0
x_5	-0.29	0.46	-0.10	0	0.57	0
x_6	-0.28	0.46	-0.09	0	0.57	0
x_7	-0.13	0.10	0.10	0	0	0
x_8	0.10	0.11	0.51	0	0	0
x_9	-0.01	0.03	-0.37	0	0	0.80
x_{10}	-0.09	0.15	0.44	0	0	0
x_{11}	0.10	0.06	0.04	0	0	0.46
x_{12}	0.09	0.01	-0.27	0	0	0
x_{13}	-0.08	0.11	-0.20	0	0	0
x_{14}	-0.09	0.13	0.47	0	0	-0.36
Variance (%)	23.62	21.16	7.87	21.21	21.10	7.41
Cumulative variance (%)	23.62	44.78	52.65	21.21	42.31	49.72

Equation (14) is the model used to generate normal data for both training and testing, Eq. (15) is the model used to generate abnormal data for testing and variables not mentioned in this equation are generated with the same model in Eq. (14).

Here, let $rand_1 \sim rand_{16}$ denote 16 different Gaussian random signals, $x_1 \sim x_{14}$ are 14 system variables. x_1, x_3 and x_4 are linear relevant. x_2, x_5 and x_6 are linear relevant. $x_7 \sim x_{12}$ are weak relevant with x_1 or x_2. x_{13} and x_{14} are irrelevant with the other 12 system variables.

In real systems, x_1, x_3, x_4 and x_2, x_5, x_6 denote two types of main features in system respectively, (i.e. in MA detection the shape features including area, perimeter and axis, another type of feature is intensity feature including the total intensity of candidate in I_{green}, average intensity of the candidate in I_{sc} etc.), $x_7 \sim x_{12}$ denote observed system variables containing large noises, although this kind of variables are relevant with the main features to some degree, actually their contributions are less important than the main features for representing the system. Besides, in some circumstances, as the noise is large, such variables may cause bad result (i.e. fault classification). x_{13} and x_{14} in the model are irrelevant with all other variables, let them denote unimportant features of the system.

In our simulation, we take 1000 observations for all variables for training, and then we get the input matrix $X \in R^{1000\times14}$. The sparsity is set at 3 which will be discussed as below. Perform the PCA and sparse PCA on the input matrix and we can obtain the corresponding load matrices respectively. Load vectors for the first three PCs and Sparse Principle Components (SPCs) are shown in Table 2. The first column of Table 2 is the first load vector by PCA. The projection of X on the first load vector has the largest variance and the projection of X on the second load vector has the second large variance and so on.

Table 2 shows the loading vectors of the first three PCs by ordinary PCA and SPCs by sparse PCA together with

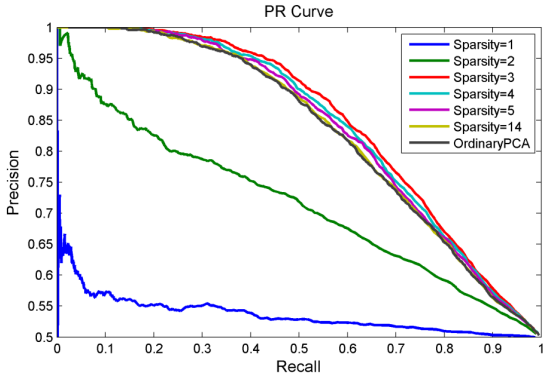


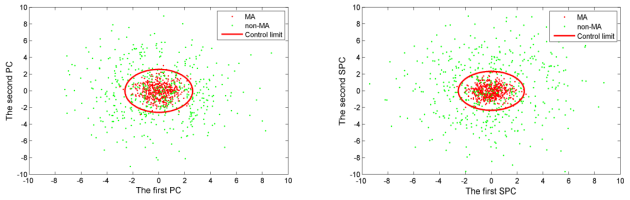
FIGURE 6. PR curves of classification results for synthetic examples.

the individual and cumulative percentage of variance in all 14 variables, accounted for by 1, 2, 3 PCs. Taking the first sparse loading vector for example, from the setting beforehand, we can know that x_1, x_3, x_4 are regarded as main features; as shown in Table 2, with the sparse loading vector, SPC1 can be represented as a liner combination of the main features. Meaning that, those irrelevant and weak relevant of features are neglected. This characteristic of sparse PCA can increase the classification accuracy which will be shown in the following PR curve.

Eq. (14) presents normal data (MA) and Eq. (15) presents abnormal data (non-MA). Here, we generate 1000 training samples using Eq. (14), and another 1500 samples are generated using Eq. (14) and Eq. (15) for validation (500 samples) and testing (1000 samples) respectively. This process is repeated for ten times. Then, we get the PR curve shown in Figure 6.

The PR curve treats precision and recall as the vertical axis and horizontal axis, respectively. According to its definition, we can learn that when the curve is closer to the top right corner, the performance of approach will be better. As can be seen from this figure, comparing with the original PCA, when sparsity equals to 3, 4 and 5, the proposed method achieves better performance than original PCA. When sparsity equals to 3, the best performance can be reached. This is coincident with the setting beforehand (x_1, x_3 and x_4 are three linear relevant features; x_2, x_5, x_6 are also three linear relevant features). When the sparsity is less than 3, classification performance decreases greatly. Thus, sparsity can be determined by PR curve.

More obviously, compare Figure 7 (a) with Figure 7(b), we can see that the scores on the first two principal components cannot separate MAs and non-MAs well, but the scores on the first two sparse principal components can do better in separating MAs and non-MAs. Actually the ellipses in Figure 7 (a) and (b) are control limits T_α^2 when α equals to 0.95. Meanwhile, judging from Figure 7, we can notice that MA scores are more compact with center on the first two sparse principal components. On the contrary, non-MA scores are more divergent, and this is an intuitive explanation of why our method can work better.



(a) (b)

FIGURE 7. Scores on the first two PCs and SPCs and the control limit with 0.95 confidence bound.

TABLE 3. The different types of images in the ROC training set.

	Resolution	Coverage of the retina	Number in training set
Type	768×576	45°	22
Type	1058×1061	45°	3
Type	1389×1383	45°	25

E. SUMMARY OF THE DETECTING PROCESS

The whole process of proposed MA detection method mainly consists of three parts. The first part is a series of preliminary works including preprocessing, candidate extraction and feature extraction. In this part, firstly, some preprocessing algorithms are applied to the green channel images for making MAs more visible. And then all the possible candidates are indentified using MSCF. Finally, a set of features are extracted for each candidate, forming the feature matrix. The second part is offline modeling, in this part, build the MA model with the proposed sparse PCA and develop the control limit L. The third part is online detecting. Given a new retinal image, firstly, repeat the process of part one, and then calculate the corresponding statistic, finally compare them with control limit L for distinguish true MAs from spurious candidates. The details of the whole detection process are described in Figure 8.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATABASE

Retinopathy Online Challenge (ROC) database includes 100 (50 train samples and 50 test samples) digital color fundus photographs which were selected from a large dataset (150,000 images) in a diabetic retinopathy screening program [26]. Since multiple screening sites use different types of cameras, basically, the fundus images were all taken with TopconNW 100, NW 200 or Canon CR5-45NM non-mydriatic cameras with three different types of Fields of View (FOV), they are presented in the dataset (type 1, type 2 and type 3, Table 3 describes all types of images) with different image resolutions.

In ROC database, only 50 training samples which are labeled as "ground truth" with locations of microaneurysm, and 50 test samples without labeling the ground truth locations. In the past, the Retina Online Challenge (ROC) organization provided a way for researchers to evaluate their methods on the test images, but now this competition website

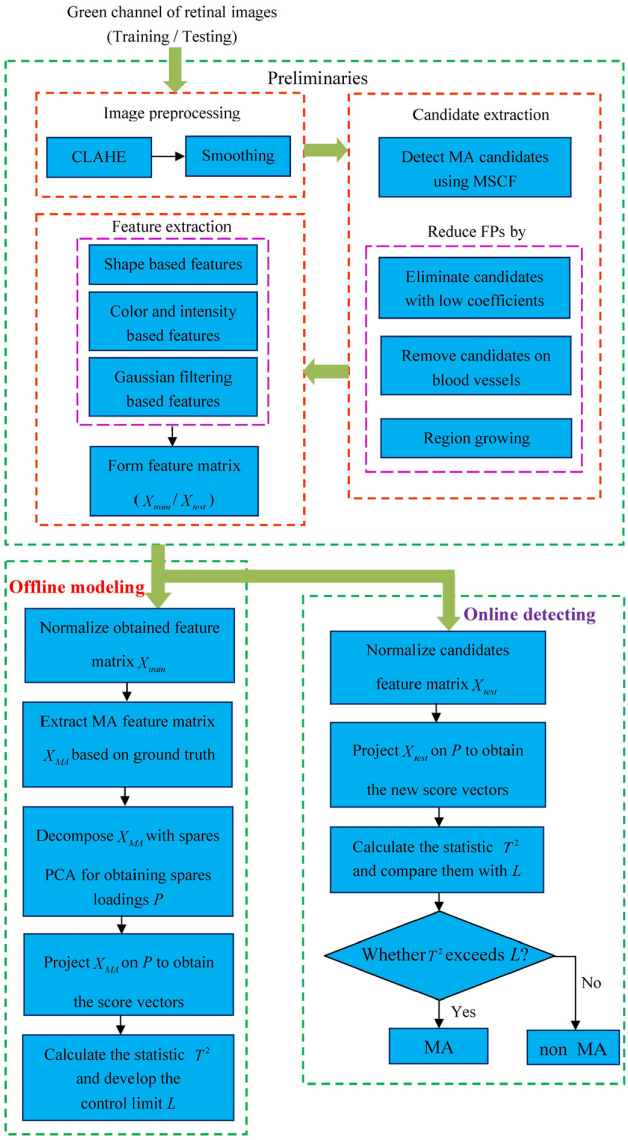


FIGURE 8. Flowchart for MA detection.

is inactive [27]. It is impossible to evaluate our method on the test images. So in our model, we employ 50 training samples to train and verify the effectiveness of our proposed method. In the training set, there are 37 digital color fundus photographs including a total of 336 microaneurysms, correspondingly no microaneurysms are identified in the remaining 13 images. In this database, 20 and 10 images are randomly selected for the training and validation sets, and the remaining 20 images are regarded as the testing set. The random sample selection is repeated 10 times and the average result is regarded as the final result. The model is established with the training set, parameters are selected via the validation set, and the performance compared with the other methods is evaluated on the testing set.

B. PARAMETER SELECTION

There is a parameter in our proposed method, sparsity which impacts the classification performance of our proposed

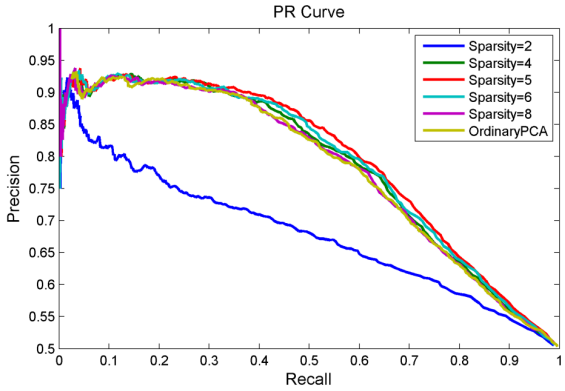


FIGURE 9. PR curves of classification results for MA detection on the validation set.

method. Since the sparsity can be determined by PR curve, which has been demonstrated in the Section 3.4.

Therefore, we employ the same manner to decide a suitable sparsity. In this experiment, 20 color fundus images are selected randomly for offline modeling and 10 images are used for parameter selection and validation. This process is repeated 10 times and the average result is regarded as our final result. As can be seen from Figure 9, when sparsity equals to 5, the proposed method can achieve the best performance on the validation set. Therefore, we set the sparsity to 5 in the following experiments. After the sparsity is determined, the threshold for T^2 statistic can be calculated with Eq. 13. Here, the value of a is set 8 by calculating CPV defined in Eq. 7.

C. EVALUATION CRITERION

Since the high number of background pixels, specificity is always close to 100% and it is not commonly used for the evaluation of MA detection methods. Therefore, Free-response Receiver Operating Characteristic (FROC) curve is adopted to evaluate the effectiveness of our proposed method in this paper. An FROC curve is the plot of sensitivity on the vertical axis and the average number of false positive per image (FPPI) on the horizontal axis. Sensitivity and FPPI are defined as follows:

$$Sensitivity = \frac{Truepositive}{Truepositive + Falsenegative}$$
$$FPPI = \frac{Falsepositive}{Totalnumberofimages} \quad (16)$$

where True positive (TP) is the number of MAs that are correctly identified; False negative (FN) is the number of incorrectly found as non-MAs; False positive (FP) is the number of incorrectly found as MAs.

The candidate extraction threshold $coeff_{AB}$ is an important parameter in MSCF, which affects the performance of algorithm. In this paper, different thresholds are employed to create the FROC curve. After an analysis of the FROC curve, the optimal value of this parameter is equivalent to 0.6.

TABLE 4. The average sensitivities and the corresponding standard deviations of different methods at various false positive points on the testing set.

	1	2	4	8	12	16	20	score
Candidate Extraction [8]	0.019±0.017	0.036±0.015	0.072±0.020	0.110±0.018	0.153±0.016	0.165±0.021	0.170±0.019	0.104±0.027
DL with SRC [13]	0.125±0.018	0.150±0.014	0.196±0.013	0.287±0.017	0.312±0.019	0.315±0.012	0.331±0.014	0.250±0.030
Math Morp [15]	0.071±0.013	0.086±0.019	0.104±0.012	0.122±0.013	0.129±0.021	0.186±0.017	0.211±0.025	0.130±0.030
FDDL [27]	0.128±0.014	0.149±0.011	0.211±0.021	0.285±0.019	0.316±0.013	0.355±0.022	0.379±0.021	0.261±0.042
Proposed method	0.135±0.012	0.155±0.010	0.232±0.015	0.288±0.016	0.325±0.011	0.370±0.019	0.420±0.012	0.275±0.024

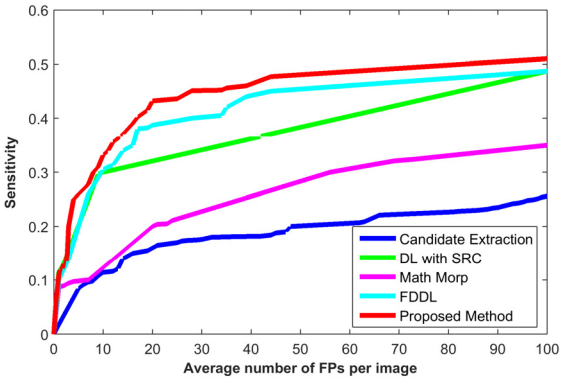


FIGURE 10. The FROC curves of the propose method with the state of the art methods using the testing set.

Using this parameter value, the candidate detection method can achieve a sensitivity of 48.31% with a FPPI of 43.13.

Figure 10 depicts the results of a comparison between the proposed method and the state-of-the-art methods [8], [13], [15], [27] using the testing set. Besides, in order to facilitate the comparison between the different methods, the score is defined as the average value of the sensitivity at seven fixed false positives per image (1, 2, 4, 8, 12, 16, and 20) on FROC curve of each method [15]. The results are shown in Table 4.

As it can be seen from Table 4, the proposed method without using any non-MA training data shows better performance and detects more MAs at the same FPPI points. Meanwhile, the score of our proposed method is 0.275, which is better than the other methods.

Lastly, we utilize the one-tailed Wilcoxon rank sum test in this paper to further demonstrate the superiority of our method to the other methods. In this experiment, the null hypothesis is that the proposed method makes no difference when compared to the other MA detection methods, and the alternative hypothesis is that the proposed method makes an improvement when compared to the other MA detection methods. For instance, if we want to compare the performance of our proposed method with that of FDDL (Proposed method vs. FDDL), the null and alternative hypotheses can be defined as H_0: $M=M_{FDDL}$ and H_1: $M>M_{FDDL}$, where M and M_{FDDL} are the medians of the sensitivity results obtained by our proposed method and FDDL. We set the significance level at 5% in our experiment. The *p*-values obtained by all pairwise Wilcoxon rank sum tests are shown in Table 5.

TABLE 5. The *p*-values of the pairwise one-tailed Wilcoxon rank sum tests on the testing set.

	The *p*-values
Our method vs. Candidate Extraction [8]	1.8267e-04
Our method vs. DL with SRC [13]	0.0046
Our method vs. Math Morp [15]	1.8267e-04
Our method vs. FDDL [27]	0.0312

From these results, we can see that they are less than 0.05, which means the null hypotheses (''medians are equal'') can be rejected in all pairwise tests and the proposed method is superior to the other methods.

V. CONCLUSION AND FUTURE WORK

Data for non-MAs vary in a large range, how to collect non-MA training set is quite subject, large training set is not only time consuming but also will cause the class imbalance problem. In this paper, an unsupervised classification method based on sparse PCA was developed to detect MA. Sparse PCA is applied to find the latent structure of MA data, once a model has been developed that reflect the MAs, any departure from standard MAs are detected by monitoring the statistics. Since there is no need to consider the non-MA class samples, the class imbalance problem can be avoided. Unsupervised classification is particular suitable for the situation that we only know the ''normal'' status (MA), ''abnormal'' status (non-MA) is unknown, or hard to obtain. Meanwhile, effective features can be selected using sparse PCA automatically. Finally, a single T^2 statistic is introduced and the control limit can be determined for distinguishing the true MAs from the spurious candidates. The experiment is carried out on ROC dataset, which shows favorable result compared with the state-of-the-art methods.

The future work includes the following issues: firstly, more effective features such as texture feature should be considered for distinguishing the MAs from FPs. Secondly, applying our proposed framework to other lesions detection such as hard exudates and hemorrhages is another interesting topic for future study.

ACKNOWLEDGMENT

The authors would like to thank the editor, the associate editor, and referees for comments and suggestions that have greatly improved this paper.

REFERENCES

[1] B. Antal and A. Hajdu, "An ensemble-based system for microaneurysm detection and diabetic retinopathy grading," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1720–1726, Jun. 2012.
[2] D. C. Klonoff and D. M. Schwartz, "An economic analysis of interventions for diabetes," *Diabetes Care*, vol. 23, no. 3, pp. 390–404, 2000.
[3] D. B. Mukamel, G. H. Bresnick, J. C. Dickinson, and D. R. Cole, "A screening approach to the surveillance of patients with diabetes for the presence of vision-threatening retinopathy," *Ophthalmology*, vol. 107, no. 1, pp. 19–24, 2000.
[4] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Rev. Biomed. Eng.*, vol. 3, no. 1, pp. 169–208, Jan. 2010.
[5] B. Antal and A. Hajdu, "Improving microaneurysm detection using an optimally selected subset of candidate extractors and preprocessing methods," *Pattern Recognit.*, vol. 45, no. 1, pp. 264–270, 2012.
[6] K. M. Adal, D. Sidibé, S. Ali, E. Chaum, T. P. Karnowski, and F. Mériaudeau, "Automated detection of microaneurysms using scale-adapted blob analysis and semi-supervised learning," *Comput. Methods Programs Biomed.*, vol. 114, no. 1, pp. 1–10, 2014.
[7] C. E. Baudoin, B. J. Lay, and J. C. Klein, "Automatic detection of microaneurysms in diabetic fluorescein angiography," *Revue D'epidemiologie Sante Publique*, vol. 32, nos. 3–4, pp. 254–261, 1984.
[8] T. Spencer, J. A. Olson, K. C. Mchardy, P. F. Sharp, and J. V. Forrester, "An image-processing strategy for the segmentation and quantification of microaneurysms in fluorescein angiograms of the ocular fundus," *Comput. Biomed. Res.*, vol. 29, no. 4, pp. 284–302, 1996.
[9] A. J. Frame *et al.*, "A comparison of computer based classification methods applied to the detection of microaneurysms in ophthalmic fluorescein angiograms," *Comput. Biol. Med.*, vol. 28, no. 3, pp. 225–238, 1998.
[10] G. G. Gardner, D. Keating, T. H. Williamson, and A. T. Elliott, "Automatic detection of diabetic retinopathy using an artificial neural network: A screening tool," *Brit. J. Ophthalmol.*, vol. 80, no. 11, pp. 940–944, 1996.
[11] G. Quellec, M. Lamard, P. M. Josselin, G. Cazuguel, B. Cochener, and C. Roux, "Optimal wavelet transform for the detection of microaneurysms in retina photographs," *IEEE Trans. Med. Imag.*, vol. 27, no. 9, pp. 1230–1241, Sep. 2008.
[12] B. Zhang, X. Wu, J. You, Q. Li, and F. Karray, "Detection of microaneurysms using multi-scale correlation coefficients," *Pattern Recognit.*, vol. 43, no. 6, pp. 2237–2248, 2010.
[13] M. Niemeijer, B. V. Ginneken, J. Staal, M. S. A. Suttorp-Schulten, and M. D. Abramoff, "Automatic detection of red lesions in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 24, no. 5, pp. 584–592, May 2005.
[14] C. I. Sánchez, R. Hornero, A. Mayo, and M. García, "Mixture model-based clustering and logistic regression for automatic detection of microaneurysms in retinal images," *Proc. SPIE*, vol. 7260, pp. 72601M-1–72601M-8, Mar. 2009.
[15] B. Zhang, F. Karray, Q. Li, and L. Zhang, "Sparse representation classifier for microaneurysm detection and retinal blood vessel extraction," *Inf. Sci.*, vol. 200, pp. 78–90, Oct. 2012.
[16] M. Niemeijer *et al.*, "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 185–195, Apr. 2010.
[17] T. Walter, P. Massin, A. Erginay, R. Ordonez, C. Jeulin, and J.-C. Klein, "Automatic detection of microaneurysms in color fundus images," *Med. Image Anal.*, vol. 11, no. 6, pp. 555–566, 2007.
[18] E. D. Pisano *et al.*, "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *J. Digit. Imag.*, vol. 11, no. 4, pp. 193–200, Nov. 1998.
[19] S. B. Júnior and D. Welfer, "Automatic detection of microaneurysms and hemorrhages in color eye fundus images," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 5, pp. 21–37, 2013.
[20] B. C. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Autom. Control*, vol. 26, no. 1, pp. 17–32, Feb. 1981.
[21] S. Valle, W. Li, and S. J. Qin, "Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods," *Ind. Eng. Chem. Res.*, vol. 38, no. 11, pp. 4389–4401, 1999.
[22] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, "A modified principal component technique based on the LASSO," *J. Comput. Graph. Statist.*, vol. 12, no. 3, pp. 531–547, 2003.
[23] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.
[24] S. J. Qin and J. Yu, "Recent developments in multivariable controller performance monitoring," *J. Process Control*, vol. 17, no. 3, pp. 221–227, 2007.
[25] S. J. Qin, "Statistical process monitoring: Basics and beyond," *J. Chemometrics*, vol. 17, nos. 8–9, pp. 480–502, 2003.
[26] M. D. Abramoff and M. S. A. Suttorp-Schulten, "Web-based screening for diabetic retinopathy in a primary care population: The eyecheck project," *Telemedicine e-Health*, vol. 11, no. 6, pp. 668–674, 2006.
[27] M. Javidi, H.-R. Pourreza, and A. Harati, "Vessel segmentation and microaneurysm detection using discriminative dictionary learning and sparse representation," *Comput. Methods Programs Biomed.*, vol. 139, pp. 93–108, Feb. 2017.

WEI ZHOU received the M.S. degree in computer science and technology from Northeast Normal University, Changchun, China, in 2015. She is currently pursuing the Ph.D. degree with Northeastern University, Shenyang, China. Her research interests include machine learning and medical imaging processing.

CHENGDONG WU is currently the Vice President of the Faculty of Robot Science and Engineering at Shenyang Northeastern University, and the Director of the Institute of Artificial Intelligence, a Professor, and a Doctoral Tutor at Northeastern University, Shenyang, China. He has long been involved in automation engineering, artificial intelligence, and teaching and researching in robot navigation. He is an expert in Chinese modern artificial intelligence and robot navigation. He is also a Special Allowance of the State Council.

DALI CHEN received the Ph.D. degree in pattern recognition and intelligent systems from Northeastern University, China, in 2008. He is currently an Associate Professor with the College of Information Science and Engineering, Northeastern University. His research interests include image processing and computer vision.

YUGEN YI was born in Jiangxi, China, in 1986. He received the B.S. degree from the College of Humanities & Sciences, Northeast Normal University, China, in 2009, the M.S. degree from the College of Computer Science and Information Technology, Northeast Normal University, in 2012, and the Ph.D. degree from the School of Mathematics and Statistics, Northeast Normal University, in 2015. He is currently a Lecturer with the School of Software, Jiangxi Normal University. His research interests include dimensionality reduction and feature extraction.

WENYOU DU received the B.S. degree in communications engineering from Northeastern University at Qinhuangdao, Qinhuangdao, China, in 2009, and the M.S. degree in control engineering from Northeastern University, Shenyang, China, in 2011, where he is currently pursuing the Ph.D. degree in control theory and control engineering. His current research interests include fault detection and diagnosis for complex industry processes.