PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2021,48(3):336~343
www.pibb.ac.cn

基于主成分分析的随机森林视网膜OCT图像分层算法研究*

李晓雯[1)] 王陆权[1)] 曾亚光[2)] 陈允照[2)] 王茗祎[2)]
钟俊平[2)] 王雪花[2)] 熊红莲[2)**] 陈 勇[1)**]
([1)] 佛山科学技术学院自动化学院，佛山 528000；[2)] 佛山科学技术学院物理与光电工程学院，佛山 528000)

摘要 视网膜是层状结构，临床上可以根据视网膜层厚度改变对一些疾病进行预测和诊断. 为了快速且准确地分割出视网膜的不同层带，本论文提出一种基于主成分分析的随机森林视网膜光学相干断层扫描技术（optical coherence tomography，OCT）图像分层算法. 该方法使用主成分分析（principal component analysis，PCA）法对随机森林采集到的特征进行重采样，保留重采样后权重大的特征信息维度，从而消除特征维度间的关联性和信息冗余. 结果表明，总特征维度在29维的情况下，保留前18维度训练速度提高了23.20%，14维度训练速度提高了42.38%，而对图像分割精度方面影响较小，实验表明该方法有效地提高了算法的效率.

关键词 光学相干断层扫描技术（OCT），视网膜分层，主成分分析，随机森林
中图分类号 R445 **DOI**：10.16476/j.pibb.2020.0270

视网膜是一种层状结构，位于眼球后壁的内层. 许多疾病可引起病患的视网膜层解剖结构变化，如糖尿病、多发性硬化、心脑血管疾病等. 因此临床上视网膜的检测是预防和诊断相关疾病的重要手段. 目前，光学相干断层扫描技术（optical coherence tomography，OCT）是最常用的眼科诊断成像技术之一，也是目前眼科疾病的研究和诊断基础之一[1]，视网膜的层状结构在OCT中表现尤为突出，由OCT仪器采集的视网膜数据为多张视网膜图像，每张图像被称为一个B-扫描，即视网膜在不同位置的切片. 图1为在中央凹处的一个B-扫描图像，其中的红色竖线为其中一个A-扫描，其边界层名如图所示. 医生观察视网膜层的细微变化来诊断疾病进程，而OCT图像视网膜层的精确分割是开展疾病诊断的重要前提和保障. 所以，基于计算机和OCT图像的自动分层算法开始兴起.

过去十几年来，关于视网膜分层的算法不断被提出，有基于图搜索[2-3]、机器学习[4]以及神经网络[5-6]等算法被使用在视网膜分层中，基于神经网络的分层方法更是近些年的热点，如Fang等[5]使用基于卷积神经网络的算法对视网膜进行分层，虽然精度高，但耗时也较长，平均一个B-扫描耗时43.1 s. 此外，也有一些基于神经网络的算法有着比传统机器学习方法更高的效率，如Zang等[6]使用神经网络和图搜索的组合算法进行视网膜分层. 但大多数神经网络方法都需要借助图形处理器（graphics processing unit，GPU）来提高效率，而在没有GPU的设备上，传统的机器学习方法则是更好的选择. 而随机森林（random forests，RF）算法误差较低、稳定性较强、算法鲁棒性高，是一种比较成熟的算法，已经被广泛地应用在医学、生物信息、生态学、遥感、网络检测等领域[4, 7-8]. 徐军[9]利用RF分割了视网膜的神经纤维层，提取视网膜的四类12种特征. Lang等[4]分割了视网膜的

* 国家自然科学基金(81601534，61771139，61805038，61705036)、国家重点研发计划(2018YFC1406601)和广东省自然科学基金(2017A030313386)资助项目.
** 通讯联系人.
熊红莲. Tel：18902331765，E-mail：yuanxiufeng138@163.com
陈勇. Tel：18029259418，E-mail：cheny@fosu.edu.cn
收稿日期：2020-07-29，接受日期：2020-09-14

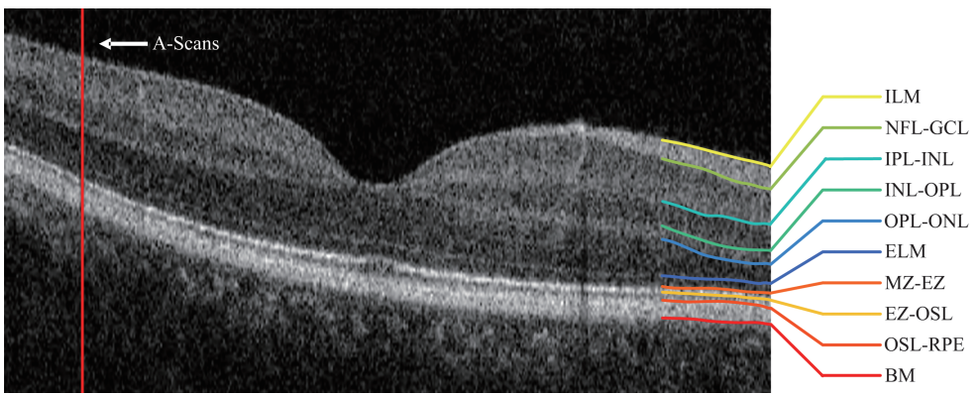

Fig. 1　B-Scan of the retina in foveal region

The full name of abbreviation on the right side of the picture are: internal limiting membrane (ILM), nerve fiber layer (NFL), ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), myoid zone (MZ), ellipsoid zone (EZ), outer segment layer (OSL), retinal pigment epithelium (RPE), Bruch's membrane (BM).

8层结构，提取了27种特征. 提取特征越多，信息量也就越多，但在使用RF算法时，选取的样本特征数量越多，算法效率也较低，虽然减少一些维度可提升效率，但也不可避免地导致信息的丢失. 主成分分析（principal component analysis，PCA）结合RF是一种有效的针对RF输入特征进行降维并且提取主要特征的方法. 此前，刘强等[10]已将此方法运用在岩体质量分类，林伟宁等[11]运用在入侵检测，黄世峰[12]将其运用在降雨量预测等.

而本文提出将利用PCA下的RF方法运用在视网膜分层上，更加快速准确地对视网膜进行分层. 主要做法如下：将提取到的所有特征维度重新映射在一个新的空间下，选取数据方差最大的维度，余下所有维度都在前一维度的垂直方向选取. 最后保留方差贡献较高的维度作为新的RF训练数据，去除方差贡献低的维度，从而降低数据维度，减少低权重信息和信息冗余，提高维度间的区分度. 新的数据使得RF分裂时得到的信息增益更高，因此能更快地构建RF分类器. 而由于保留信息比例大，所以在精度方面影响较小. 最终提高了RF分层算法的效率.

1　理论分析算法

1.1　PCA

PCA被用在许多领域中，如图像处理、数据降维、股市预测等[8, 13-14]，它能将众多具有一定相关性的指标重新组合（重采样），提取重要的维度，形成一组新的相互无关（正交）的综合指标来代替原指标，从而消除冗余信息. 方法如下，以一个去中心化的二维特征$X=\{a_i,b_i\}^T$为例，a_i,b_i为不同的特征维度，X的协方差矩阵为：

$$Cov(X)=\frac{1}{m}XX^T=\begin{bmatrix}\frac{1}{m}\sum_{i=1}^{m}a_i^2 & \frac{1}{m}\sum_{i=1}^{m}a_ib_i\\ \frac{1}{m}\sum_{i=1}^{m}a_ib_i & \frac{1}{m}\sum_{i=1}^{m}b_i^2\end{bmatrix}\tag{1}$$

矩阵中的$\sum_{i=1}^{m}a_i^2$为特征的方差，代表特征值；$\sum_{i=1}^{m}a_ib_i$为特征之间的协方差，其大小代表特征之间的相关性. 为了让特征维度之间相互无关，需要各特征之间相互正交，即对角化（协方差为0）. 而通过求解公式（1），可得到一组特征向量，将特征向量按照其所对应的特征值λ由大到小进行重排，并取前k个向量（$k<n$），就得到了空间映射矩阵P. 假设λ中的特征值$\lambda_1<\lambda_2<\cdots<\lambda_n$，则保留前$k$维所包含的原数据信息比例$\eta$，可通过公式（2）求得：

$$\eta=\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i}\tag{2}$$

P可将X映射到新的空间Y中，此时Y便是特征相互正交的数据[13]. 这样就将后面信息量少的高维特征去除了，同时达到降低特征关联和消除信息冗余的目的.

1.2　RF

RF是一个非常成熟的多决策树分类算法，它使用随机的维度选取和有放回地抽样，使得生成的决策树更具随机性，从而保证算法的准确度和鲁棒性[7]. 但在特征之间的区分度（独立性）较低时，

RF分类器的训练效率会降低，其根本原因是：特征间的区分度低时，决策树得到的信息增益较小，分类变得困难，从而需要更多的分支进行分类判断，信息增益公式为：

$$g(D|A) = H(D) - H(D|A) \quad (3)$$

式中为$g(D|A)$信息增益值，$H(D)$为分支前的总体熵值，$H(D|A)$为在A条件下预分支后的熵值[15]. 当每次分裂时样本特征之间的区分度越高，信息增益就越大，能更快地完成决策树的构建，因此提高特征间的区分度是提高RF决策树构建效率的一个方法.

1.3　基于PCA的RF

RF视网膜分层算法任务中需要大量的数据样本，样本量的削减可能会导致分层的结果不佳，因此实验在特征维度上着手以提高RF算法的效率. 为了提高效率，RF算法希望得到有区分度的信息，而PCA恰好给予了这一条件. 因此实验使用PCA对RF视网膜分层中的特征维度进行降维，以减少冗余的特征信息，提高样本特征的纯度. 由于RF算法分为训练、预测两部分，视网膜数据也需要分为训练和测试两部分数据进行处理，本文算法中训练和预测数据处理如图2a所示.

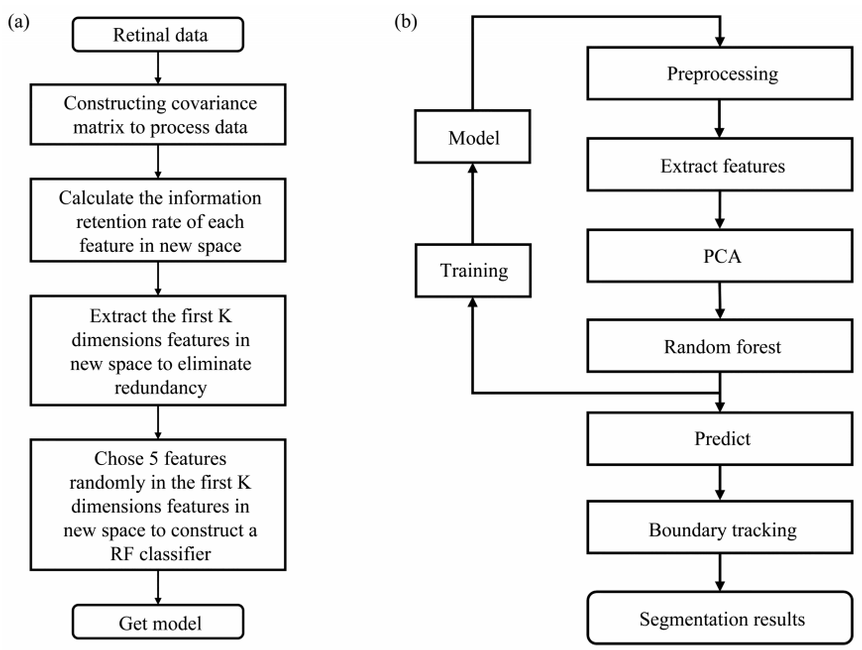

Fig. 2　The training processing of PCA−RF algorithm and the general process of this paper

(a) The training and prediction data processing diagram. (b) The flow chart of algorithm.

2　视网膜OCT图像分层算法

分层算法主要分为图像预处理、特征提取、PCA降维以及随机分类器训练和预测4大部分，本文算法流程如图2b所示. 首先，对获取的眼底视网膜数据进行归一化、压平等预处理，使图像更利于分层. 之后，再对视网膜图像进行特征选取，并采用PCA进行特征维度降维. 同时为训练样本的边界进行人工标注. PCA降维后的特征和人工标注随后被输入到RF分类器中进行学习以生成模型，生成的模型便可对新的样本进行预测分层.

2.1　视网膜数据预处理

一份视网膜实验数据集一共包括128张视网膜图像. 视网膜图像的预处理主要分为归一化和压平处理两部分[2, 4]，归一化可使各视网膜图像在数值更接近，提高算法的计算速度，更利于分类器学习. 本文算法采用的是min-max归一化，公式（4）所示：

$$I'(m,n) = \left(I(m,n) - I_{min}\right) / \left(I_{max} - I_{min}\right) \quad (4)$$

受扫描角度及眼底结构影响，采集的视网膜轮廓是弯曲的，会导致边界追踪时出错. 为此要对视网膜数据进行压平，首先选择能比较快被找到的边界作为基准线，本文选择BM作为基准线，因为BM的梯度较大，波动不大，且不易与其他边界混淆，而MZ-EZ虽然梯度大，但容易找错，与距离较近的OSL-RPE产生混淆. 压平时，计算在每个

A-扫描（图1）中BM边界与整体BM平均高度之间的偏移量，以此为依据移动每个A-扫描，使整个视网膜体的BM处在同一平面上.

2.2　特征选取与PCA降维

实验提取视网膜数据体的29个特征以供RF分类器进行学习，比Lang论文中的27个特征多了两个梯度特征[4]. 具体类别和提取方式如下：

a. 邻域类别特征. 当前像素在当前B-扫描的3×3邻域，共9个邻域特征（3×3邻域）.

b. 梯度类别特征. 原图在当前像素A-扫描方向上的一阶梯度和二阶梯度（2个），高斯各向异性滤波在度（−10°、0°、10°）与像素尺度σ（x_1，y_1）=（5，1）和σ（x_2，y_2）=（10，2）的全组合（6张图像）的A-扫描方向上一阶和二阶梯度值（12个）.

c. 环境类别特征. 当前像素分别向下15、25、35个像素的11×11域在A-扫描方向上的梯度平均值，共3个.

d. 位置类别特征. 当前像素相对于视网膜中央凹的x、y、z 3个维度距离，共3个.

选取的特征维度越多，包含的信息就越多，但信息冗余和运算量也随之增加. 为在减少运算时间的同时保留更多的信息，实验使用PCA对数据的特征维度进行空间映射，该方法降低各维度之间的相关性并去除信息含量低的维度. 表1中记录了降维后取前K个特征维度时，所保留的信息比例. 这些比例是由10个视网膜数据训练集的平均值所构成. 由表中数据可知，降维后仅保留前8个维度，信息保留率就达到了99%，这也证明PCA确实能有效地减少信息冗余.

Table 1　The retention rate when maintaining first *K* dimensions after dimensionality reduction

Dimensions	Retention rate of information/%
2	88.16
4	96.24
6	98.03
8	99.02
10	99.81
12	99.92
14	99.98

2.3　分类器训练和边界追踪

本实验中RF的参数以及训练样本的数量是参考Lang的文章中最小集参数（minimal set of parameters，MSP）所设置的[4]. 在RF分类器的训练中，实验的训练集随机选取了两个视网膜数据体，每个数据体中随机选取12张B-扫描（共24张）作为训练样本，并对其边界进行人工标记. RF的决策树数量设置为20，随机特征数设置为5. 随后将降维后新选取的特征维度和人工标签输入分类器中以生成模型. 值得注意的是，在用模型预测测试样本前，要将测试数据映射到以训练集转换矩阵为基矩阵的空间中进行降维，保证其是在同一空间下的数据处理. 预测时，RF中的树会根据训练时学习到的信息对B-扫描的每个像素点进行投票，而非得到具体的边界位置，即RF得到的是以原图为大小的图上对每一边界在所有像素点上树的投票数目.

为了更加细化地分割边界，我们对RF得到的概率图进行边界追踪. 首先将RF得到的投票转换为0~1之间的概率. 以森林有100棵树为例，对于某一像素点，80棵树为它投ILM类别票，则该点在ILM边界概率图中的值为0.8. 下图为某张B-扫描在保留不同维度时NFL-GCL边界的概率图（图3a，b）. 得到某个边界的概率图后，根据Canny边界探测算法的原理，使用非最大抑制法找到图中每个A-扫描里最大的概率点作为边界点，保留这些点作为该B-扫描的某条边界[4]. 在找到一条边界后，需要对下一条边界限定搜索区域以避免边界的交叉. 最后使用3次样条插值对这些点进行曲线拟合，得到最终边界，图3（c、d）中红线分别为对a、b概率图边界进行边界追踪的结果.

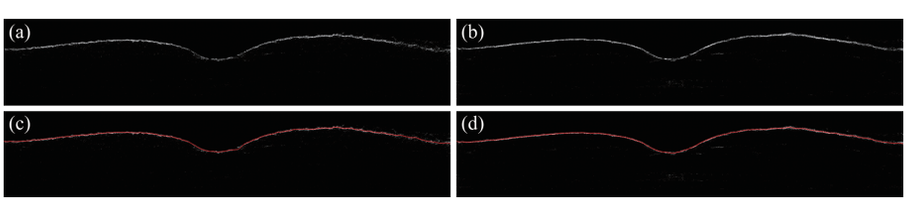

Fig. 3　The probability graph of NFL−GCL in a B−scan and its boundary tracking results when keeping different dimension

(a) Probability graph predicted by 18 dimensions feature model which using PCA. (b) Probability graph predicted by 29 dimensions feature model. Red lines in c, d are boundary tracking results of a, b respectively.

3　结果与讨论

3.1　实验数据和环境

本实验使用inter Core（TM）i5-8300H CPU平台搭配16G的运行内存作为硬件平台，使用拓普康的眼底断层扫描系统进行数据采集，共采集40个正常眼底视网膜数据体. 每只眼睛为一个视网膜数据体，其中有128张B-扫描，每张B-扫描中有512列A-扫描，每列A-扫描中有885个像素，数据体在实际中的大小为6×6×2.3 mm^3，其中每个B-扫描和A-扫描之间的实际间距分别为46.8 μm和11.7 μm，A-扫描中每个像素的实际间距为2.6 μm.

为了验证方法的有效性，实验从40个样本中选取了10个作为训练集，剩余30个作为测试集. 每一次从训练集中随机选取2个样本，每个样本随机选12张B-扫描进行人工标注并训练，中央凹的和非中央凹的B-扫描各占一半. RF的决策树数量设置为20，随机特征数设置为5. 对于同一组训练样本，需要记录分类器在不同维度下生成模型的时长，并计算单个测试样本最终分层结果和随机12张人工标签之间的平均绝对误差（*MAE*）.

3.2　PCA降维前后算法表现

测试结果的数据比对主要分为2个部分，一是降维前后的训练时间变化，二是降维前后分层精度的变化. 虽然降维后保留前8个特征维度就能涵盖原数据99%的信息，但实际测试时，要保留10个以上的维度才有较为可观的分层结果. 此外，由于降维对预测速度的影响较小，因此本文仅讨论降维对训练速度的影响. 不同维度的情况下耗费的平均训练时长如图4a所示，随着保留的维度减少，训练的时间也减少了，从原29维降到了14维后，训练时间降低了将近一半（42.38%）. 由此可见，PCA降维确实提升了视网膜数据在RF中的训练速度.

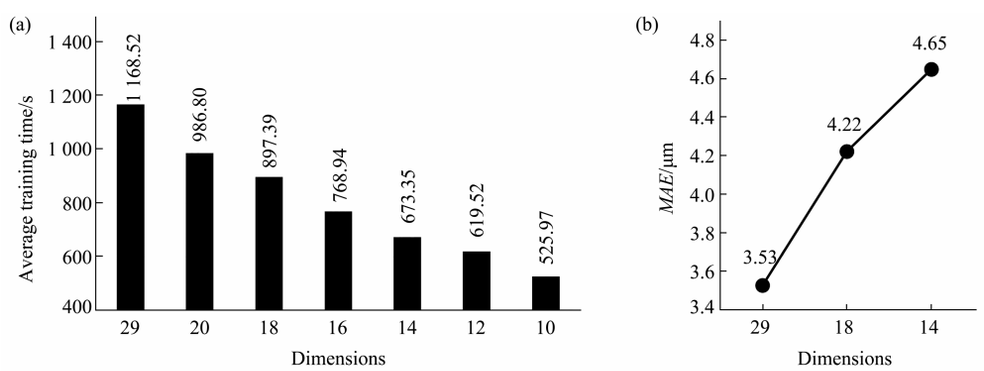

Fig. 4　Training time and segmentation accuracy analysis before and after dimensionality reduction

(a) The average training time when keeping different number of dimensions. (b) The whole 10 boundaries' *MAE*, which were between ground truth and retaining 14, 18, 29 dimensions.

而在误差方面，本实验统计了不同维度下分层结果与人工标注之间的*MAE*. 结果显示，在保留的维数降低后，分层的精度也出现了不同程度地下降. 在权衡了训练速度及误差之后，实验选取了保留14和18维的结果，与原29维度的结果进行对比，分别比较10个边界的整体*MAE*，以及精确到各个边界的*MAE*（图4b，图5）. 从速度和误差数据上看，保留14维度的结果在比原维度的训练时间少了将近一半的情况下，*MAE*提高1.1 μm，而保留18维度的时间虽减少得不如14维度的多，但*MAE*仅比原维度提高0.7 μm，这个结果在去除了多个维度的情况下是可以接受的.

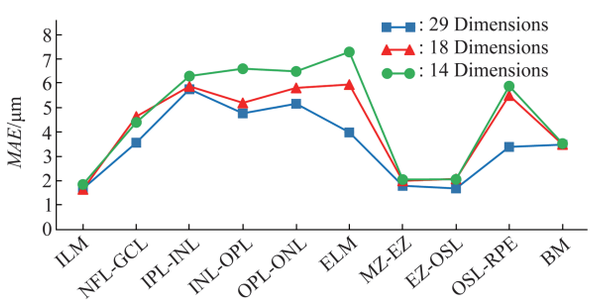

Fig. 5　The *MAE* of each boundary between the ground truth and the final segmentation which were using model trained by 14，18 and 29 dimensions features

保留14、18维和原维度的分层结果如图6所示，由上至下分别为14、18和原29维度的分层结果. 从分层效果来讲，在保留18维的情况下，虽然降维后在某些边界上的波动较大，但大部分边界仅有较小的波动，对分层结果的影响不大. 而仅保留14维的结果相对较差，但胜在速度快. 从整体上看，14维能大致分出每层位置；而18维与原维度的结果更为相近，有着不错的准确度，这一点在图5中也有所体现，除ELM和OSL-RPE等边界外，其他边界的*MAE*变化不大. 血管阴影区域的分层效果如图7所示，可见在降维后，该位置的分层结果于原维度结果也非常相近，与整体分层结果分析的情况一致，没有受到大的影响. 上述结果说明PCA+RF方法在视网膜分层中确实起到了提高效率的作用，并且在分层结果上也有着较不错的准确率，证明了此方法的有效性.

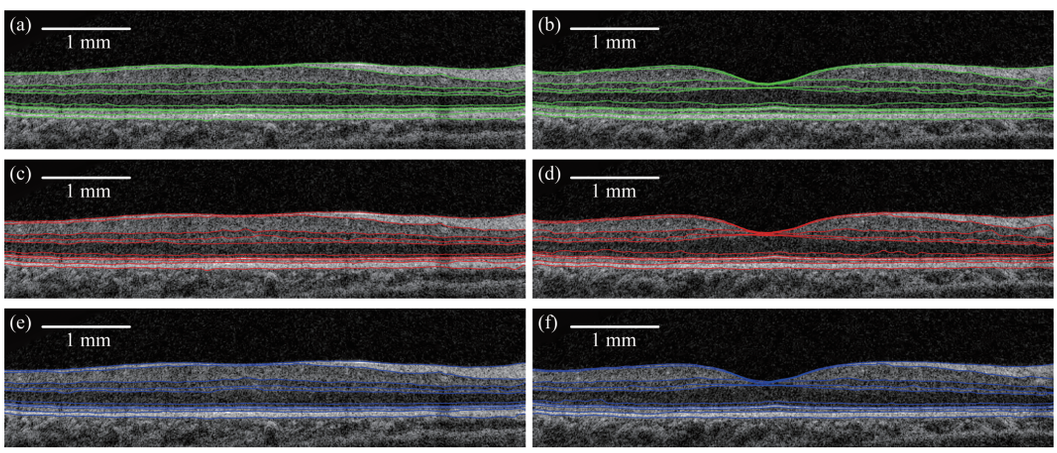

Fig. 6　The result of segmentation which were using model trained by 14，18 and 29 dimensions features

Green lines in (a, b) are segmentation results which predicted by 14 dimensions features model, red lines in (c, d) are results for 18 dimensions features and blue lines in (e, f) are results for original 29 dimensions features. Images on the left are results of B-scans in parafoveal region, and images on the right are results of B-scans in foveal region.

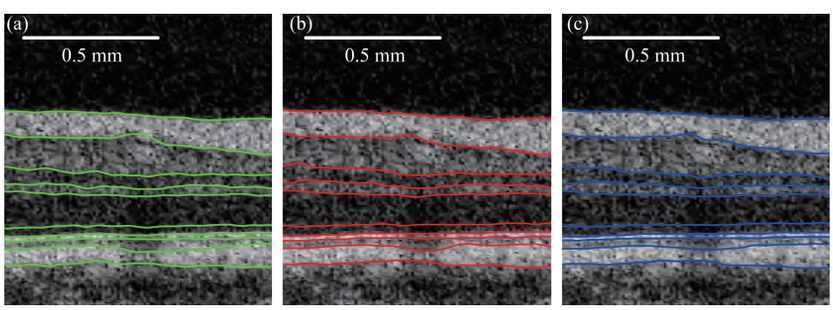

Fig.7　Segmentation performance in vessel shadow area

From left to right are the segmentation results of vessel shadow region predicted by 14, 18, and 29 dimensions models.

4　结　　论

本文在RF视网膜分层任务中，使用PCA降低特征维度的方法，解决了特征维度之间关联性大和信息冗余多的问题. PCA降维后特征维度的纯度更高，使得RF决策树在生成时的信息熵下降更快，从而提升了分类器的训练效率，减少了训练的时长，为加快RF视网膜分层的训练速度提供了一种新的研究思路和手段.

参　考　文　献

[1] Huang D, Swanson E A, Lin C P, *et al*. Optical coherence tomography. Science, 1991, **254**(5035): 1178-1181

[2] Chiu S J, Li X T, Nicholas P, *et al*. Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation. Opt Express, 2010, **18**(18): 19413-19428

[3] Zhang Z, Tang Y, Zeng X, *et al*. Projection area evaluation of macular edema by optical coherence tomography images with automatic retinal segmentation. Progress in Biochemistry and

Biophysics, 2020, **47**(10): 1097-1107

[4] Lang A, Carass A, Hauser M, *et al*. Retinal layer segmentation of macular OCT images using boundary classification. Biomed Opt Express, 2013, **4**(7): 1133-1152

[5] Fang L, Cunefare D, Wang C, *et al*. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. Biomed Opt Express, 2017, **8**(5): 2732-2744

[6] Zang P, Wang J, Hormel T T, *et al*. Automated segmentation of peripapillary retinal boundaries in OCT combining a convolutional neural network and a multi-weights graph search. Biomed Opt Express, 2019, **10**(8): 4340-4352

[7] Breiman L. Random forests. Machine Learning, 2001, **45**(1): 5-32

[8] 余忠永, 黄俊, 许二敏, 等. 基于PCA降维的多特征行人再识别. 信息通信, 2019, **196**(4): 13-16
Yu Z, Huang J, Xu E, *et al*. Information & Communications, 2019, **196**(4): 13-16

[9] 徐军. SD-OCT视网膜图像神经纤维层的检测与研究[D]. 江苏:南京理工大学, 2017
Xu J. Detection and Study of Nerve Fiber Layer in SD-OCT Retinal Image[D]. Jiangsu: Nanjing University of Science and Technology, 2017

[10] 刘强, 李夕兵, 梁伟章. 岩体质量分类的PCA-RF模型及应用. 黄金科学技术, 2018, **26**(1): 49-55
Liu Q, Li X, Liang W. Gold Science and Technology, 2018, **26**(1): 49-55

[11] 林伟宁, 陈明志, 詹云清, 等. 一种基于PCA和RF分类的入侵检测算法研究. 信息网络安全, 2017, **203**(11): 50-54
Lin W, Chen M, Zhan Y, *et al*. Netinfo Security, 2017, **203**(11): 50-54

[12] 黄世锋. 基于主成分分析和随机森林的短时降雨量预测. 电子世界, 2019 (1): 22-23
Huang S. Electronics World, 2019 (1): 22-23

[13] 谢心蕊, 雷秀仁, 赵岩. MI和改进PCA的降维算法在股价预测中的应用. 计算机工程与应用, 2019, **56**(21): 139-144
Xie X, Lei X, Zhao Y. Computer Engineering and Applications, 2019, **56**(21): 139-144

[14] Rodarmel C, Shan J. Principal component analysis for hyperspectral image classification. Surveying and Land Information Science, 2002, **62**(2): 115-122

[15] 汪桂金. 随机森林算法的优化改进及其并行化研究[D]. 江西: 南昌大学, 2019
Wang G. Optimization and Improvement of Random Forest Algorithm and Its Parallelization[D]. Jiangxi: Nanchang University, 2019

Random Forest Retinal Segmentation in OCT Images Based on Principal Component Analysis*

LI Xiao-Wen[1)], WANG Lu-Quan[1)], ZENG Ya-Guang[2)], CHEN Yun-Zhao[2)], WANG Ming-Yi[2)], ZHONG Jun-Ping[2)], WANG Xue-Hua[2)], XIONG Hong-Lian[2)**], CHEN Yong[1)**]

([1)]*Automatic College, Foshan University, Foshan* 528000, China;

[2)]*School of Physics and Optoelectronic Engineering, Foshan University, Foshan* 528000, China)

Abstract　The retina is a layered structure, and some diseases can be clinically predicted and diagnosed based on the change in the thickness of the retinal layer. To segment the different layers of the retina quickly and accurately, this study proposes a random forest algorithm based on principal component analysis (PCA). The algorithm uses PCA to resample the normalized features collected from the retinal images and retains the feature information dimensions with significant weight, thereby eliminating the relevance between the different feature dimensions and information redundancy. After PCA, the number of features can be reduced obviously, but still retains 99% information. Random forests algorithm applies the features to learn and predict the location of retinal layer boundaries. We extract each pixels values of retinal boundaries, producing an accurate probability map for each boundary. Experimental results show that when the total number of feature dimensions decreased from 29 to 18, the training speed of the model increased by 23.20%. By contrast, when the number of feature dimensions was 14, the training speed increased by 42.38%. However, the effect on image segmentation accuracy was not obvious. Thus, it is found that this method effectively improves the efficiency of the algorithm.

Key words　optical coherence tomography (OCT), retina segment, principal component analysis (PCA), random forest (RF)

DOI：10.16476/j.pibb.2020.0270

* This work was supported by grants from The National Natural Science Foundation of China（81601534，61771139，61805038，61705036），National Key R&D Program of China（2018YFC1406601）and Natural Science Foundation of Guangdong Province（2017A030313386）.

**Corresponding author.

XIONG Hong-Lian. Tel：86-18902331765，E-mail：yuanxiufeng138@163.com

CHEN Yong. Tel：86-18029259418，E-mail：cheny@fosu.edu.cn

Received：July 29，2020　Accepted：September 14，2020