Computerized Macular Pathology Diagnosis in Spectral Domain Optical Coherence Tomography Scans Based on Multiscale Texture and Shape Features

Yu-Ying Liu,[1] *Hiroshi Ishikawa,*[2,3] *Mei Chen,*[4] *Gadi Wollstein,*[2] *Jay S. Duker,*[5] *James G. Fujimoto,*[6] *Joel S. Schuman,*[2,3] *and James M. Rehg*[1]

Purpose. To develop an automated method to identify the normal macula and three macular pathologies (macular hole [MH], macular edema [ME], and age-related macular degeneration [AMD]) from the fovea-centered cross sections in three-dimensional (3D) spectral-domain optical coherence tomography (SD-OCT) images.

Methods. A sample of SD-OCT macular scans (macular cube 200 × 200 or 512 × 128 scan protocol; Cirrus HD-OCT; Carl Zeiss Meditec, Inc., Dublin, CA) was obtained from healthy subjects and subjects with MH, ME, and/or AMD (dataset for development: 326 scans from 136 subjects [193 eyes], and dataset for testing: 131 scans from 37 subjects [58 eyes]). A fovea-centered cross-sectional slice for each of the SD-OCT images was encoded using spatially distributed multiscale texture and shape features. Three ophthalmologists labeled each fovea-centered slice independently, and the majority opinion for each pathology was used as the ground truth. Machine learning algorithms were used to identify the discriminative features automatically. Two-class support vector machine classifiers were trained to identify the presence of normal macula and each of the three pathologies separately. The area under the receiver operating characteristic curve (AUC) was calculated to assess the performance.

Results. The cross-validation AUC result on the development dataset was 0.976, 0.931, 0939, and 0.938, and the AUC result on the holdout testing set was 0.978, 0.969, 0.941, and 0.975, for identifying normal macula, MH, ME, and AMD, respectively.

Conclusions. The proposed automated data-driven method successfully identified various macular pathologies (all AUC > 0.94). This method may effectively identify the discriminative features without relying on a potentially error-prone segmentation module. (*Invest Ophthalmol Vis Sci.* 2011;52:8316–8322) DOI:10.1167/iovs.10-7012

Spectral-domain optical coherence tomography (SD-OCT) is a noncontact, noninvasive, three-dimensional (3D) imaging technique that performs optical sectioning at micrometer resolution. It is widely used in ophthalmology for identifying the presence of disease and its progression.[1] This technology measures the optical back-scattering of the tissues, making it possible to visualize intraocular structures and diagnose ocular diseases, such as glaucoma and macular hole, objectively, and quantitatively.

Although OCT imaging technology continues to evolve, technology for automated OCT image analysis and interpretation has not kept pace. With OCT data being generated at increasingly larger amount and higher sampling rates, there is a strong need for automated analysis to support disease diagnosis and tracking. This need is further amplified by the fact that an ophthalmologist making a diagnosis under standard clinical conditions does not have the assistance of a specialist in interpreting OCT data beforehand. A software system that is capable of automated interpretation can potentially assist clinicians in making clinical decisions efficiently in busy daily routines.

To our knowledge, there has been no prior work on automated macular pathology identification in OCT images, with the goal of directly predicting the presence probability for each macular pathology in a given cross-sectional frame; this automated method can be helpful to support disease diagnosis especially in situations where qualified readers are not easily accessible.

Automated pathology identification in ocular OCT images is complicated by three factors. First, the co-existence of pathologies with other pathologic changes (e.g., epiretinal membrane, vitreous hemorrhage) can complicate the overall appearance, making it challenging to model each pathology individually. Second, there is high appearance variability within each pathology (e.g., in macular hole cases, the holes can have different widths, depths, and shapes, and some can be covered by incompletely detached tissues, making explicit pathology modeling difficult). Third, the measurement of re-

From the [1]School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia; the [2]UPMC Eye Center, Eye and Ear Institute, Ophthalmology and Visual Science Research Center, Department of Ophthalmology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania; the [3]Department of Bioengineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania; the [4]Intel Science and Technology Center, Carnegie Mellon University, Pittsburgh, Pennsylvania; the [5]New England Eye Center, Tufts Medical Center, Tufts University School of Medicine, Boston, Massachusetts; and the [6]Department of Electrical Engineering and Computer Science and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Presented at the annual meeting of the Association for Research in Vision and Ophthalmology, Fort Lauderdale, Florida, May 2010, and at the International Conference on Medical Image Computing and Computer Assisted Intervention, Beijing, China, September 2010.

Supported by National Institutes of Health Grants NIH R01-EY013178, R01-EY011289, and P30-EY008098 (Bethesda, MD); The Eye and Ear Foundation (Pittsburgh, PA); Research to Prevent Blindness, Inc. (New York, NY); and Intel Labs, Intel Corporation (Mountain View CA).

Submitted for publication December 7, 2010; revised April 29 and September 6, 2011; accepted September 6, 2011.

Disclosure: **Y.-Y Liu**, None; **H. Ishikawa**, None; **M. Chen**, None; **G. Wollstein**, None; Carl Zeiss Meditec (F), Optovue (F); **J.S. Duker**, Topcon Medical Systems (F), Alcon (C), Genentech (C), Ophthotech (C), Paloma Pharmaceuticals (C); **J.G. Fujimoto**, Optovue (C, I); **J.S. Schuman**, None; **J.M. Rehg**, None

Corresponding author: Hiroshi Ishikawa, UPMC Eye Center, Eye and Ear Institute, 203 Lothrop Street, Room 835.2, Pittsburgh, PA 15213; ishikawah@upmc.edu.

flectivity of the tissue is affected by the optical properties of the overlying tissues (e.g., opaque media in the vitreous area or blood vessels around retinal surfaces will block or absorb much of the transmitted light respectively, and thus produce shadowing effects). As a result of these factors, attempts to hand craft a set of features or rules to identify each pathology are unlikely to generalize well. Instead, direct encoding of the statistical distribution of low-level image features and training discriminative classifiers based on a large expert-labeled dataset may achieve a more robust performance.

In this study, a machine learning–based method for automatically identifying the presence of pathologies from a fovea-centered cross section in a macular SD-OCT scan was developed. Specifically, the presence of the normal macula (NM) and each of the three macular pathologies macular hole (MH), macular edema (ME), and age-related degeneration (AMD) were identified separately in the cross section through the foveal center. This single-frame–based method can serve as a basic component for examining the complete set of frames from the volume.

In this work, the automated software that makes diagnostic suggestions is solely based on the interpretation of image appearances, so as to serve as a stand-alone component for OCT image interpretation. Note that for a true clinical diagnosis, all the available information (e.g., the results of OCT image analysis in conjunction with other ancillary tests) would be considered together to make the final diagnostic decision.

A preliminary version of this work was presented in our prior paper.[2] This article significantly extends our previous publication in several areas: improved automated method, detailed labeling agreement analysis by three ophthalmologists, new ground truth based on majority opinion and complete consensus, evaluation of our original and new method, and several additional experiments, such as effect of training set size, effect of data with inconsistent labeling, and performance on a separate testing dataset collected after the method development stage, which is representative of future unseen data.

Methods

Subjects and Image Acquisition

The study subjects were enrolled at the University of Pittsburgh Medical Center Eye Center or at the New England Eye Center. All subjects had comprehensive ophthalmic examination followed by SD-OCT macular cube scan (Cirrus HD-OCT; Carl Zeiss Meditec, Dublin, CA). The training dataset (dataset A), consisting of 326 macular SD-OCT scans from 136 subjects (193 eyes), was used for deriving the best algorithmic and parameter settings by cross-validation. The testing dataset (dataset B), containing another 131 macular SD-OCT scans from 37 subjects (58 eyes) collected after the method development stage, was used for testing the performance on novel images.

Since the OCT manufacturer's recommended signal strength (SS) is 8 or above in 1 to 10 scale, all our enrolled images were qualified SS ≥ 8 criteria. The original scan density was either 200 × 200 × 1024 or 512 × 128 × 1024 samplings in 6 × 6 × 2-mm volumes. All horizontal cross section images were rescaled to 200 × 200 for computational efficiency. For each of the scans, the horizontal cross section through the foveal center was then manually selected by one expert ophthalmologist, and this image served as the basis for analysis in this study.

The study was approved by the Institutional Review Board committees of the University of Pittsburgh and Tufts Medical Center (Boston, MA) and adhered to the Declaration of Helsinki and Health Insurance Portability and Accountability Act regulations, with informed consent obtained from all subjects.

Subjective Classification of Images

A group of OCT experts masked to any clinical information independently identified the presence or absence of normal macula and each of MH, ME and AMD in the fovea-centered frame. Note that a combination of pathologies can coexist in one cross section. For the MH category, both macular hole and macular pseudohole were included to simplify the discrimination of all holelike structures from the other cases. Dedicated labeling software was developed where only the preselected fovea-centered frame was presented in a randomized order.

For dataset A, three OCT experts gave the pathology labels for each scan, and the majority opinion of the three experts was identified for each pathology and used as the "ground truth" in our method development stage. For dataset B, two of the three experts provided the labels for each scan. For each pathology, the scans with consistent labels were selected for performance evaluation, and the scans with different labels were excluded.

Automated Classification Method

Our automated method encodes the appearance properties of the retinal images directly, by constructing a global image descriptor based on spatially distributed, multiscale texture, and shape features, combined with machine learning techniques to automatically learn the classifiers for identifying each pathology from a large expert labeled training set. This method does not rely on a retinal segmentation

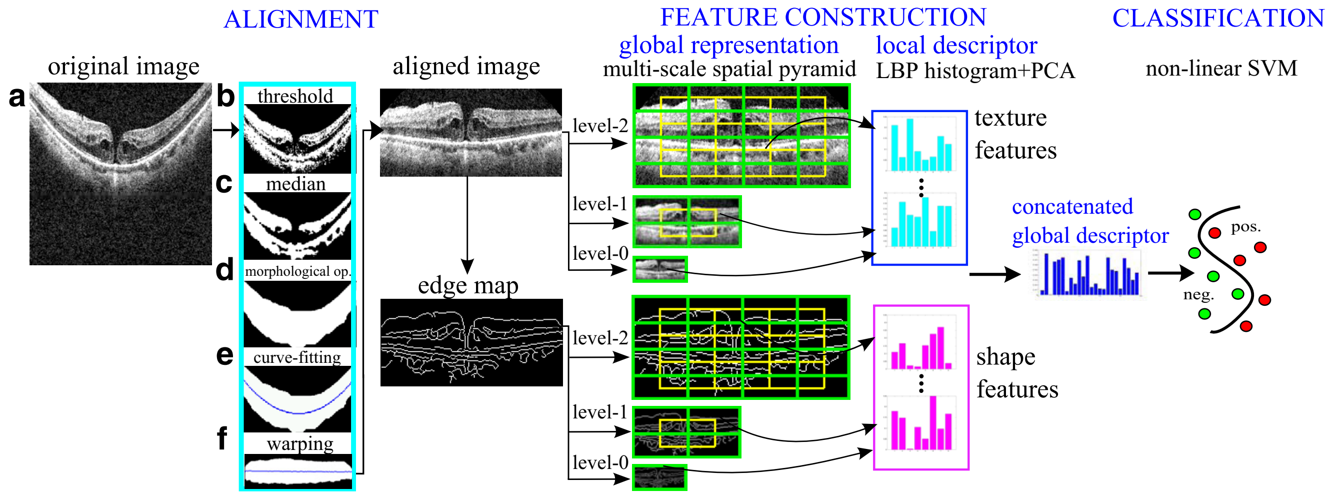

Figure 1. Stages of our approach. Morphologic op., morphologic operations; LBP, local binary patterns; PCA, principle component analysis; SVM, support vector machine.

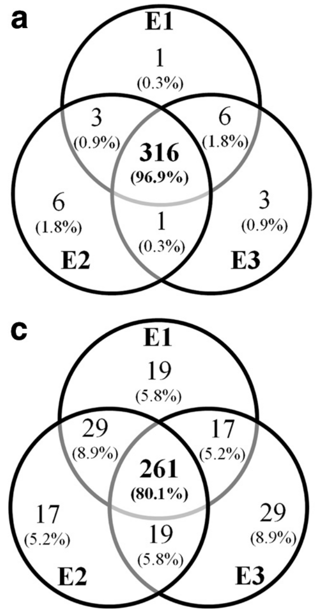

FIGURE 2. Venn diagram of labeling agreement among the three experts on all macular categories in database A. The actual scan numbers and the percentage are both shown. E1, E2, E3, the three experts.

module to extract the features and thus avoids a major source of analysis failure.

An earlier formulation of our automated method, which has been published,[2] uses only texture features. In this study we extended and enhanced the approach by incorporating the shape property of the retinal images in addition to the texture. In brief, our method consists of three main steps as illustrated in Figure 1. First, image alignment is performed to remove the curvature and center the image to reduce the appearance variation across scans. Second, a global image descriptor is constructed from the aligned image and its derived Canny edge image.[3] Multiscale spatial pyramid (MSSP)[4] is used as the global representation for capturing the spatial organization of the retina in multiple scales and spatial granularities. To encode each local block, the dimension-reduced local binary pattern (LBP) histogram[5] based on principle component analysis (PCA) is used as the local block descriptor. The local features derived from each spatial block in the multiple rescaled images and their edge images are concatenated in a fixed order to form the overall global descriptor. These histogram-based local features are used to encode the texture and shape characteristics of the retinal image, respectively. Finally, for each pathology, a two-class, nonlinear support vector machine (SVM)[6] classifier with radial basis function (RBF) kernel and estimated probability values is trained using the image descriptors and their labels from the training set.

For detailed information in validating each component (MSSP, LBP) of our approach, please refer to our prior study.[2]

Experimental Settings

In developing the method on dataset A, 10-fold cross validation was used at the subject level where all images from the same subject were put together in either the training or testing set in each fold. Note that by cross validation, each testing fold is novel with respect to its corresponding training fold. In our training phase, both the original image and its horizontal flip are used as the training instances for enriching our training set. After running all 10-fold training and testing, the 10-fold testing results are aggregated and the area under the receiver operator characteristic curve (AUC) is computed. To get a more reliable assessment of the performance, 6 different random 10-fold data splitting were generated, and the above procedure is run for each of the six splits. The mean and SD of the 6 AUCs were reported as the performance metric on the developing set.

To test the statistical significance of performance difference between any two algorithmic settings, DeLong et al.[7] test was adopted to compare the two receiver operating characteristics (ROC) curves. If under DeLong test, one setting is better than the other ($P \leq 0.05$) for all six different data splits, then the performance of the former setting is claimed to be better; otherwise, the difference in the performance of the two settings is declared not to be significant.

TABLE 1. Kappa Values for Dataset A

κ	NM	MH	ME	AMD
E1, E2	0.94	0.92	0.76	0.76
E1, E3	0.97	0.78	0.69	0.73
E2, E3	0.93	0.76	0.71	0.77

E1, E2, and E3 represent the three experts. The values within 0.60–0.80 represent substantial but imperfect agreements.

TABLE 2. Positive Scans, Eyes, and Subjects from Dataset A

Statistics	NM	MH	ME	AMD
Scan	81	74	203	74
Eye	66	36	116	37
Subject	65	33	90	26

Data are from a total of 326 scans of 136 subjects (193 eyes), as defined by the majority opinion of the three experts. Note that each scan was labeled with coinciding macular findings.

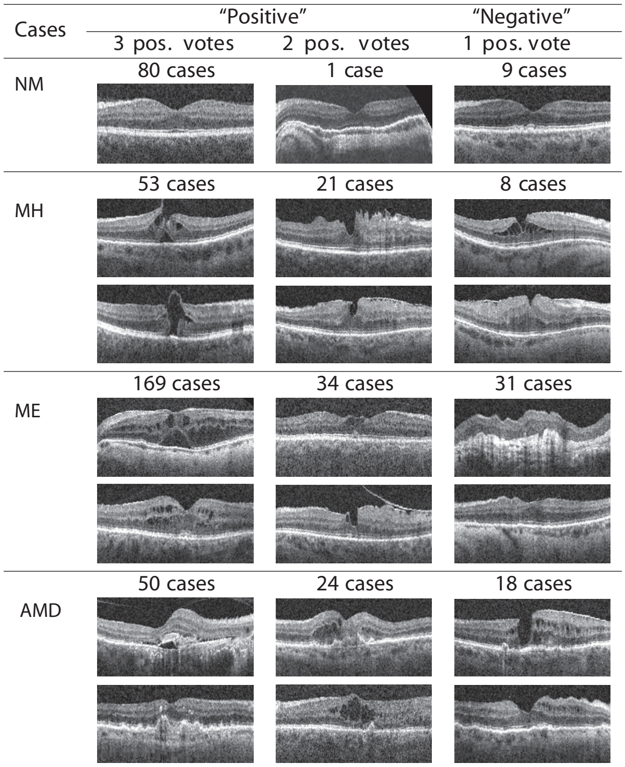

FIGURE 3. The number of cases and representative examples in database A where all three experts, two experts, or only one expert gave "positive" labels for the presence of normal macula and each pathology. Note that images in the first two columns were defined as positive while the ones in the last column were regarded as negative in our majority-opinion–based ground truth. The images without total agreement usually contain early pathologies that were subtle and occupied small areas.

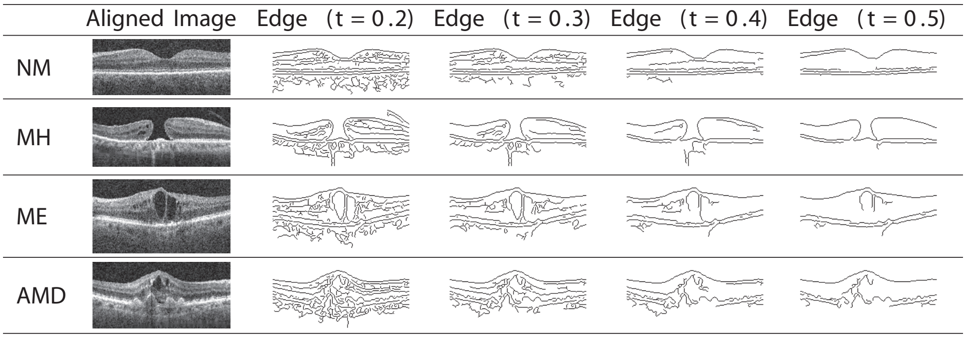

FIGURE 4. Examples of the aligned retinal images and their Canny edge maps derived under different edge-detection thresholds *t* for each macular category. The smaller the value of *t*, the more edges are retained.

After performing detailed analysis on dataset A, the best algorithmic settings and parameters determined for identifying each pathology are then applied to the test dataset B. The performance on dataset B is thus representative for the generalization ability of the proposed approach. (for MH category, unfortunately, dataset B did not contain macular hole cases, which coincides with real clinical situations, since macular hole has low occurrence (approximately 3.3 cases in 1000 in those persons older than 55 years).[8] To deal with this situation, for MH performance testing only, the training and testing dataset was reorganized such that 80% of MH cases originated from dataset A were randomly sampled and included in the training set and the rest were included in the testing set).

RESULTS

Interexpert Labeling Agreement and the Ground Truth on Dataset A

The labeling agreement among the three experts on dataset A was illustrated in Figure 2 using the Venn diagram. The complete agreement among the experts for NM, MH, ME, and AMD was 96.9%, 91.1%, 80.1%, and 87.1%, respectively. The κ statistic was calculated to assess the pair-wise experts' labeling agreement, as shown in Table 1. All κ values for identification of normal macula were high (all $\kappa > 0.93$) and for MH, the κ value from one expert pair (experts 1 and 2) was high (0.92). However, all κ values for ME and AMD were within the 0.61 to 0.80 range, which represented substantial but imperfect agreement.

The majority opinion of the image labeling was used as the ground truth so that the standard would not be biased toward any specific expert. The number of images for each macular category as defined by the ground truth is shown in Table 2.

To further assess how many positive and negative cases in our ground truth result from inconsistent labeling, in Figure 3, the statistics and several representative examples were shown for each pathology where all three experts, only two experts, or just one expert gave the "positive" label. Note that the images labeled as positive by only one expert were treated as "negative" cases in our ground truth. It was found that the quantity of images having only one positive vote is considerable for ME and AMD (31 and 18 cases, respectively), revealing larger ambiguity in their identification.

Performance of Automated Classification Method on Dataset A

Different feature settings: texture (T) alone, shape (S) alone, and in combination (TS) were tested on dataset A, so that the discriminative power of each feature type for each pathology can be evaluated. For shape features, the edge detection threshold, denoted as *t*, was tested at various values so that different quantities of edges were obtained and encoded (Fig. 4). The AUCs for the different feature settings were reported in Table 3. The best AUCs for NM, MH, ME, and AMD, were 0.976, 0.931, 0.939, and 0.938, derived from the setting: TS ($t = 0.4$), S ($t = 0.4$), TS ($t = 0.4$), and TS ($t = 0.2$), respectively; their ROC curves generated from one of six random data splits were shown in Figure 5.

Regarding the edge detection thresholds *t* for shape features, it was discovered that for NM, ME, and AMD, the AUC results under different *t* settings were all within 1% in AUC; but for MH, the performance is much more sensitive to the choice of *t* (the AUC was 0.888, 0.901, 0.931, and 0.911 when *t* varied from 0.2 to 0.5) with the best performance at t =0.4; this suggests that for MH, encoding the stronger edges is more helpful in identifying the hole structures; the weaker edges ($t = 0.2$) may add noise instead and distract the classifiers.

The statistical significance of the performance difference under different feature settings was also evaluated. It was found that for NM, TS outperformed T, although the absolute gain is small (0.7% in AUC); thus, including shape features can provide additional useful information. For MH, S is significantly better than using T and TS, with a large AUC difference (8.5%) between S and T; this reveals that using shape feature alone is sufficient to capture the distinct contours of MH. For ME, T and TS was significantly better than using S (1.6% AUC difference), but TS and T has similar performance; this suggests that en-

TABLE 3. AUC Results from Dataset A of Texture Features, Shape Features, and Their Combinations under the Best Edge Detection Threshold *t*

AUC	Texture (T)	Shape (S)	Texture + Shape (TS)	Significance Test
NM	0.969 ± 0.002	0.971 ± 0.002 (t=0.4)	**0.976 ± 0.002** (t=0.4)	T ≈ S, **T < TS,** S ≈ TS
MH	0.846 ± 0.011	**0.931 ± 0.005** (t=0.4)	0.919 ± 0.005 (t=0.4)	**T < S, T < TS,** S ≈ TS
ME	**0.939 ± 0.004**	0.923 ± 0.005 (t=0.3)	**0.939 ± 0.004 (t=0.4)**	**S < T,** T ≈ TS, **S < TS**
AMD	0.925 ± 0.008	0.931 ± 0.005 (t=0.2)	**0.938 ± 0.006** (t=0.2)	T ≈ S ≈ TS
Average	0.908 ± 0.008	0.932 ± 0.005	**0.936 ± 0.006**	

Data are the mean ± SD. The best results for each macular category are shown in bold. The rightmost column shows the results of the DeLong test at a $P = 0.05$ significance level, where < shows that the setting on the right performs better than that on the left, and ≈ represents no significant difference.

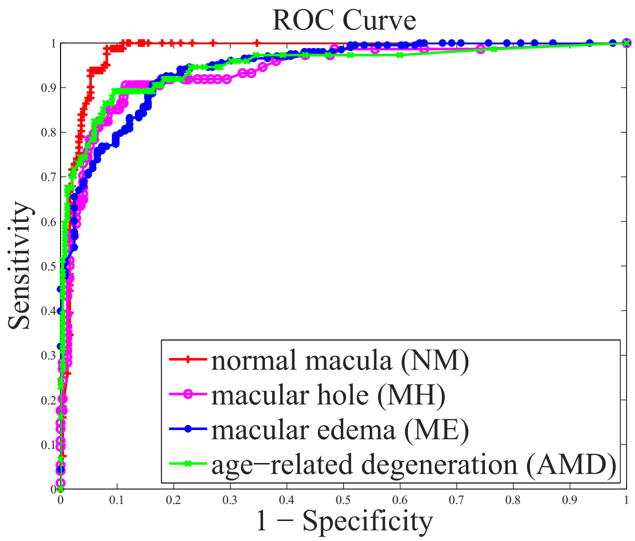

Figure 5. ROC curve of one run of 10-fold cross-validation on all images in dataset A. The best feature setting for each macular pathology was used. Feature setting: TS (t = 0.4), S (t = 0.4), TS (t = 0.4), and TS (t = 0.2) for NM, MH, ME, and AMD, respectively.

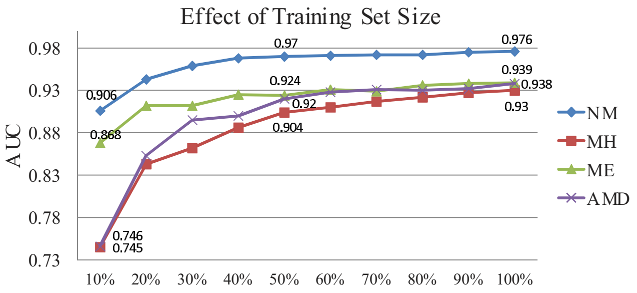

Figure 6. AUC results with respect to a varied training set size from dataset A. For each training fold, 10%, 20%, . . . , 100% of the positive and negative subjects were sampled and used for training, whereas the testing fold was unchanged. Feature setting: TS (t = 0.4), S (t = 0.4), TS (t = 0.4), and TS (t = 0.2) for NM, MH, ME, and AMD, respectively.

coding the intensity patterns (textures) is more informative than just describing the shapes. For AMD, all three feature settings (T, S, TS) had no significant difference, but using combined features (TS) achieved the best AUC performance, suggesting that both feature types are useful.

In implementation, for NM, ME, and AMD, the feature vectors are computed from the aligned retinal image directly, which is 200 pixels in width; for MH, the features are extracted from the further down-sized image (rescaled to 100 pixels in width). This rescaling for MH improves the performance by 3% consistently under different feature type settings and suggests that removing the details or noise residing in the original resolution can help identify the hole structures.

Performance Comparison between Experts and Automated Method on Dataset A

To compare the labeling performance of the automated method to that of each expert against the majority-based ground truth, the balanced accuracy (average of sensitivity and specificity) of the automated method and each expert was computed. For the automated method, the best balanced accuracy was derived from the ROC curve. The results were detailed in Table 4. Overall, the automated analysis method achieved good balanced accuracy for NM (95.5%), but relatively lower performance for MH, ME, and AMD (89.7%, 87.3%, and 89.3%). The automated software was inferior to the experts in most cases, but when compared to expert 3, the performance differences were all within 5% for all categories (the difference is −3.9%, +3.2%, −4.4%, and −2.7% for NM, MH, ME, and AMD, respectively).

Performance Using Varied Training Set Size on Dataset A

The AUC performances of the automated method with respect to varied training size on dataset A were also studied. The 10-fold cross-validation setting was still used, but now for each training fold, k% of positive and negative subjects were sampled and used for training, whereas the testing fold remained the same. The results with settings of k = 10, 20, . . . , 100 are plotted in Figure 6. The AUC results of 10%, 50%, and 100% training set were 0.906. 0.970, and 0.976 for NM; 0.745, 0.904, and 0.931 for MH; 0.868, 0.924, and 0.939 for ME, and 0.746, 0.920, and 0.938 for AMD. These results show that using more training data can improve performance in all categories. For MH, a larger gain (2.7%) and clearer increasing trend from 50% to 100% is observed, suggesting that adding more training instances for MH can improve the performance the most.

From the theoretical viewpoint, using more training data are always desirable for learning-based approaches, since this can help discover the true discriminative information from more representative images, mitigate the overfitting problems, and thus achieve better generalization performance.

Performance Using Only Images with Complete Consensus on Dataset A

To understand the influence of cases where there is inconsistent labeling, we conducted an experiment using only images with complete labeling agreement for each pathology separately. In this setting, 316 (96.9%), 297 (91.1%), 261 (80.1%), and 284 (87.1%) images from the original 326 images were selected for NM, MH, ME, and AMD identification, respectively

Table 4. Balanced Accuracy of Each of the Three Experts and the Automated Method against the Majority-Opinion-Based Ground Truth on Database A

Accuracy	Expert 1	Expert 2	Expert 3	Automated
NM	**99.8** (100, 99.6)	98.4 (98.8, 98.0)	99.4 (100, 98.8)	95.5 (99.4, 91.5)
MH	**99.4** (100, 98.8)	98.3 (98.6, 98.0)	86.5 (73.0, 100)	89.7 (89.1, 90.3)
ME	92.4 (99.5, 85.4)	**94.9** (94.6, 95.1)	91.7 (89.2, 94.3)	87.3 (87.5, 87.0)
AMD	**94.2** (93.2, 95.2)	94.0 (89.2, 98.8)	92.0 (85.1, 98.8)	89.3 (89.7, 88.8)
Average	**96.5** (98.2, 94.8)	96.4 (95.3, 97.5)	92.4 (86.8, 98.0)	90.5 (91.4, 89.4)

Shown is the balanced accuracy (sensitivity, specificity). For the automated method, the best feature setting for each pathology was adopted (TS, t = 0.4; S, t = 0.4; TS, t = 0.4; and TS, t = 0.2, for NM, MH, ME, and AMD, respectively) The best balanced accuracy was derived from the mean of the output of the six runs.

Table 5. AUC Results of Using All of Dataset A (326 Images) in Comparison with Those Obtained with Only the Images of Complete Consensus from the Three Experts for Each Pathology Separately

AUC on Dataset A	NM	MH	ME	AMD
All images (326 scans)	0.976	0.931	0.939	0.938
Images of complete consensus from the three experts	0.984	0.932	0.985	0.968

Included were 316 (96.9%), 297 (91.1%), 261 (80.1%), and 284 (87.1%) images for NM, MH, ME and AMD, respectively.

(as illustrated in the Venn diagram in Fig. 3). The AUC results are shown in Table 5.

It was found that when using only images with complete consensus, the performance for NM and MH is slightly enhanced (~1%), but it is much better for ME (from 0.939 to 0.985) and AMD (from 0.938 to 0.968). This suggests that the larger ambiguity in ME and AMD identification, as noted in their lower κ values, is indeed a major factor in influencing the performance of the automated method.

Performance on the Separate Testing Dataset B

To test the performance on the holdout dataset B, the pathology classifiers were trained using the images from dataset A, with the best algorithmic settings determined in analyzing dataset A (TS, $t = 0.4$; S, $t = 0.4$; TS, $t = 0.4$; and TS, $t = 0.2$ for NM, MH, ME, and AMD, respectively). For this experiment, the ground truth was defined by the consensus between the same two experts for both datasets. The consensus includes 96.9%, 95.4%, 88.0%, and 90.5% of 326 scans from dataset A for training and 94.7%, 100%, 90.0%, and 84.7% of 131 scans from dataset B for testing, for NM, MH, ME, and AMD, respectively. The pathology distribution for both datasets is detailed in Table 6. The AUC result and the ROC curve are shown in Table 7 and Figure 7, respectively.

The AUC is 0.978, 0.969, 0.941, and 0.975, and the best balanced accuracy is 95.5%, 97.3%, 90.5%, and 95.2% for NM, MH, ME, and AMD, respectively. The AUC performance on all pathologies are good (AUC $>$ 0.94) and comparable to the cross-validation AUC results on the training dataset A (AUC $>$ 0.93). Our results suggest that the proposed method is effective in identifying pathologies for future unseen images.

Table 6. Number of Positive Scans, Eyes, and Subjects versus the Total Cases, as Defined by the Consensus of the Two Experts on Dataset A (Training) and B (Testing)

	NM	MH*	ME	AMD
Training statistics				
Scan	80/316	49/287	190/287	59/295
Eye	66/187	27/176	109/180	27/178
Subject	65/133	26/128	84/130	21/133
Testing statistics				
Scan	22/124	21/153	59/118	81/111
Eye	13/54	8/66	29/54	31/50
Subject	10/36	6/43	23/34	20/33

The consensus includes 96.9%, 95.4%, 88.0%, and 90.5% of 326 scans from dataset A, and 94.7%, 100%, 90.0%, and 84.7% of 131 scans in dataset B. In each cell, the number of positive cases versus the total cases is shown.

* For the MH category, unfortunately, dataset B did not contain MH cases, which coincides with real clinical situations, since MH has a very low prevalence rate. Therefore, for MH diagnosis performance testing only, the training and testing datasets were organized in a way that 80% of MH cases originating from dataset A were randomly sampled and included in the training set, and the rest (six subjects) were included in the testing set.

Table 7. Testing Performance on Dataset B, Based on the Pathology Classifiers Trained on Dataset A

Performance on Dataset B	NM	MH	ME	AMD
AUC	0.978	0.969	0.941	0.975
Best balanced accuracy, %	95.5	97.3	90.5	95.2

The ground truth for this experiment was defined by the consensus from the two experts for both datasets. The consensus includes 96.9%, 95.4%, 88.0%, and 90.5% of 326 scans from dataset A, and 94.7%, 100%, 90.0%, and 84.7% of 131 scans from dataset B, for NM, MH, ME, and AMD, respectively. The number of positive cases versus total cases is shown.

Discussion

In this study, a machine-learning–based approach was proposed to identify the presence of normal macula and several macular pathologies—MH, ME, and AMD—from a fovea-centered cross section in a macular SD-OCT scan. To our knowledge, this study is the first to automatically classify OCT images for various macular pathologies. A large dataset (dataset A) containing 326 scans from 136 subjects with healthy macula or assorted macular pathologies was used for developing the methodology, and a separate dataset (dataset B), with 131 scans from 37 subjects, was used as a holdout testing set.

On the developing dataset (dataset A), the performance of our automated analysis achieved $>$0.93 AUC results for all macular pathologies using 10-fold cross-validation, with particularly good performance on identifying the normal macula (AUC = 0.976). This can be attributed to the reduced variation in normal appearance across scans. For pathology identification, the performance decreased somewhat, probably due to the greater within-category appearance varia-

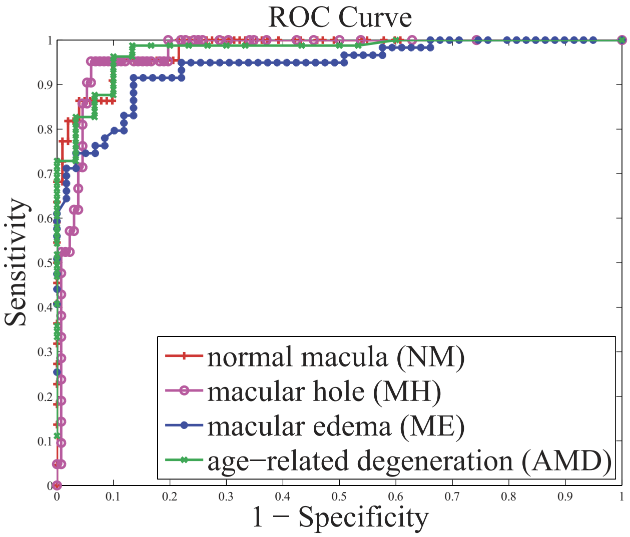

Figure 7. ROC curve of testing on dataset B, based on the pathology classifiers trained using images from dataset A. The ground truth for this experiment was defined by the consensus of the two experts (experts 1 and 2) on both datasets. The statistics of pathology distribution are shown in Table 6. The feature and parameter setting for each pathology was determined using dataset A only. Feature setting: TS ($t = 0.4$), S ($t = 0.4$), TS ($t = 0.4$), and TS ($t = 0.2$) for NM, MH, ME, and AMD, respectively.

tions, lack of sufficient training data, especially for MH and AMD, and the ambiguity existing in the majority-opinion–based ground truth, as shown in the κ agreement analysis between the experts.

By analyzing the performance on dataset A, we were able to study the discriminative power of using texture or shape features alone and in combination. It was found that, with the DeLong test for MH, shape features were more effective than texture features, whereas for ME, texture features outperformed shapes. This makes sense, since MHs are marked by the distinct contours of holes, while detection of ME requires intensity-comparison information (e.g., dark cystic areas embedded in the lighter retinal layers). For NM and AMD, the combined features achieved the highest AUC results, but this setting did not significantly outperform using either feature alone. However, it is possible that when a larger training set is available, using all complimentary features can result in superior performance, since the overfitting phenomenon in the high-dimensional feature space can be mitigated, and the true discriminative information can be more effectively represented.

The AUC results with respect to varied training set size (10%, 20%, . . . , 100%) on dataset A were also presented. It was discovered that exploiting more training data can consistently enhance the performance in all categories, especially for MH. Training on additional MH cases can boost the performance the most.

To understand the influence of inconsistent labeling in our majority-opinion–based ground truth from dataset A, the AUC results from using only images with complete consensus for each pathology were also presented. The much higher AUC results for ME and AMD (0.985 and 0.968, respectively) suggest that our current method is more effective when the two classes (presence and absence) can be well separated. However, in reality, there are always subtle cases residing in the gray area in between, causing ambiguity in dichotomy labeling. One possible future direction is to use refined labeling (e.g., by pathologic degree: absent, early, or advanced), and to explore whether this setting can result in improved labeling consistency and superior performance in automated software. This new methodology may demand a larger amount of training data in discriminating different pathologic stages.

Our method achieved good AUC results ($>$0.94 for all pathology categories) on the hold-out testing set (dataset B), when using images from the developing dataset (dataset A) for classifier training. This performance is promising in classifying future unseen images.

The proposed method has several advantages. First, our histogram-based image features directly capture the statistical distribution of appearance characteristics, resulting in objective measurements and straightforward implementation. Second, our method is not limited to any one pathology and can be applied to identify additional pathologies. Third, the same approach can be used to examine other cross sections besides the foveal slice, as long as the labeled cross sections from the desired anatomic location are also collected for training.

The limitation of the present study is that only the fovea-centered frame for each 3D SD-OCT scan was analyzed, and that frame was manually selected. In practice, every slice in the 3D scan data should be examined so that any abnormality can be identified, even when no pathology is observed at the fovea-centered frame (an unlikely event). This study is designed as a foundation for extending to analyzing each slice in the volume.

In our future work, the present slice diagnosis method will be extended to analyze each slice in the entire cube, once the pathology labeling for each cross section can be gathered. The most straightforward way is to train a set of y-location indexed pathology classifiers using the labeled slice set from the same quantized y location relative to the fovea. By using location-specific classifiers, the normal and abnormal anatomic structures around similar y locations can be modeled more accurately, and the entire volume can be examined. Once the eye motion artifacts in the macular scans can be reliably corrected, the efficacy of volumetric features will be investigated for pathology identification. An automated method for fovea localization is also desirable so that the entire process is fully automated.

In conclusion, an effective approach was proposed to computerize diagnosis of multiple macular pathologies in retinal OCT images. Our results (AUC $>$ 0.94) demonstrate that the proposed spatially distributed multiscale texture and shape descriptors combined with a data-driven framework can effectively identify the discriminative features without relying on a potentially error-prone segmentation module. Our method may provide clinically useful tools to support disease identification, improving the efficiency of OCT-based examination.

References

1. Schuman JS. Spectral domain optical coherence tomography for glaucoma. *Trans Am Ophthalmol Soc.* 2008;106:426–458.
2. Liu Y-Y, Chen M, Ishikawa H, Wollstein G, Schuman J, Rehg JM. Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid with local binary patterns. *International Conference on Medical Image Computing and Computer Assisted Intervention.* 2010;6361:1-9.
3. Canny J. A Computational Approach To Edge Detection. *IEEE Trans Pattern Analysis and Machine Intelligence.* 1986;8:679–698.
4. Wu J, Rehg JM. *Where Am I: Place Instance and Category Recognition Using Spatial PACT.* Presented at IEEE Computer Vision and Pattern Recognition 2008, Anchorage, Alaska, June 2008; New York: IEEE/Wiley; 2008;1–8.
5. Ojala T, Pietikäinen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell.* 2002;24:971-987.
6. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM TIST.* 2011;2(3):27.1–27.27.
7. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44:837–845.
8. Luckie A, Heriot W. Macular hole: pathogenesis, natural history, and surgical outcomes. *Aust N Z J Ophthalmol.* 1995;23:93-100.