Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/168320

Please be advised that this information was generated on 2018-07-08 and may be subject to change.

Biomedical Optics EXPRESS

Automatic differentiation of color fundus images containing drusen or exudates using a contextual spatial pyramid approach

Mark J. J. P. van Grinsven,[1,*] Thomas Theelen,[2] Leonard Witkamp,[3] Job van der Heijden,[3] Johannes P. H. van de Ven,[2] Carel B. Hoyng,[2] Bram van Ginneken,[1] Clara I. Sánchez[1,2]

[1]*Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands*
[2]*Department of Ophthalmology, Radboud University Medical Center, Nijmegen, The Netherlands*
[3] *KSYOS TeleMedical Center, Amstelveen, The Netherlands*
[*]*Mark.vanGrinsven@radboudumc.nl*

Abstract: We developed an automatic system to identify and differentiate color fundus images containing no lesions, drusen or exudates. Drusen and exudates are lesions with a bright appearance, associated with age-related macular degeneration and diabetic retinopathy, respectively. The system consists of three lesion detectors operating at pixel-level, combining their outputs using spatial pooling and classification with a random forest classifier. System performance was compared with ratings of two independent human observers using human-expert annotations as reference. Kappa agreements of 0.89, 0.97 and 0.92 and accuracies of 0.93, 0.98 and 0.95 were obtained for the system and observers, respectively.

© 2016 Optical Society of America

OCIS codes: (100.0100) Image processing; (100.2960) Image analysis; (100.5010) Pattern recognition.

References and links

1. J. W. Y. Yau, S. L. Rogers, R. Kawasaki, E. L. Lamoureux, J. W. Kowalski, T. Bek, S.-J. Chen, J. M. Dekker, A. Fletcher, J. Grauslund, S. Haffner, R. F. Hamman, M. K. Ikram, T. Kayama, B. E. K. Klein, R. Klein, S. Krishnaiah, K. Mayurasakorn, J. P. O'Hare, T. J. Orchard, M. Porta, M. Rema, M. S. Roy, T. Sharma, J. Shaw, H. Taylor, J. M. Tielsch, R. Varma, J. J. Wang, N. Wang, S. West, L. Xu, M. Yasuda, X. Zhang, P. Mitchell, T. Y. Wong, and Meta-Analysis for Eye Disease (META-EYE) Study Group, "Global prevalence and major risk factors of diabetic retinopathy," Diabetes Care **35**(3), 556–564 (2012).
2. S. Sivaprasad, B. Gupta, R. Crosby-Nwaobi, and J. Evans, "Prevalence of diabetic retinopathy in various ethnic groups: a worldwide perspective,"Surv. Ophthalmol. **57**(4), 347–370 (2012).
3. X. Zhang, J. B. Saaddine, C.-F. Chou, M. F. Cotch, Y. J. Cheng, L. S. Geiss, E. W. Gregg, A. L. Albright, B. E. K. Klein, and R. Klein, "Prevalence of diabetic retinopathy in the United States, 2005-2008," JAMA **304**(6), 649–656 (2010).
4. W. L. Wong, X. Su, X. Li, C. M. G. Cheung, R. Klein, C.-Y. Cheng, and T. Y. Wong, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," Lancet Glob. Health **2**(2), e106–e116 (2014).
5. D. S. Friedman, B. J. O'Colmain, B. Muñoz, S. C. Tomany, C. McCarty, P. T. V. M. de Jong, B. Nemesure, P. Mitchell, J. Kempen, and Eye Diseases Prevalence Research Group, "Prevalence of age-related macular degeneration in the United States," Arch. Ophthalmol. **122**(4), 564–572 (2004).
6. N. Cheung, P. Mitchell, and T. Y. Wong, "Diabetic retinopathy," Lancet **376**(9735), 124–136 (2010).

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016

7. J. Karnon, C. Czoski-Murray, K. Smith, C. Brand, U. Chakravarthy, S. Davis, N. Bansback, C. Beverley, A. Bird, S. Harding, I. Chisholm, and Y. C. Yang, "A preliminary model-based assessment of the cost-utility of a screening programme for early age-related macular degeneration," Health Technol. Assess. **12**(27), iii-iv, ix–124 (2008).
8. D. K. Nagi, C. Gosden, C. Walton, P. H. Winocour, B. Turner, R. Williams, J. James, and R. I. G. Holt, "A national survey of the current state of screening services for diabetic retinopathy: ABCD-Diabetes UK survey of specialist diabetes services 2006," Diabet. Med. **26**(12), 1301–1305 (2009).
9. M. J. J. P. van Grinsven, Y. T. E. Lechanteur, J. P. H. van de Ven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Automatic drusen quantification and risk assessment of age-related macular degeneration on color fundus images," Invest. Ophthalmol. Vis. Sci. **54**(4), 3019–3027 (2013).
10. R. T. Smith, M. A. Sohrab, N. Pumariega, Y. Chen, J. Chen, N. Lee, and A. Laine, "Dynamic soft drusen remodelling in age-related macular degeneration," Br. J. Ophthalmol. **94**(12), 1618–1623 (2010).
11. A. D. Mora, P. M. Vieira, A. Manivannan, and J. M. Fonseca, "Automated drusen detection in retinal images using analytical modelling algorithms,"Biomed. Eng. Online. **10**(1), 59 (2011).
12. C. I. Sánchez, M. García, A. Mayo, M. I. López and R. Hornero, "Retinal image analysis based on mixture models to detect hard exudates," Med. Imgage. Anal. **13**(4), 650–658 (2008).
13. U. M. Akram and S. A. Khan, "Automated detection of dark and bright lesions in retinal images for early detection of diabetic retinopathy," J. Med. Syst. **36**(5), 3151–3162 (2012).
14. M. Niemeijer, B. van Ginneken, S. R. Russel, M. S. A. Suttorp-Schulten, and M. D. Abràmoff, "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis," Invest. Ophthalmol. Vis. Sci. **48**(5), 2260–2267 (2007).
15. C. Agurto, E. S. Barriga, V. Murray, S. Nemeth, R. Crammer, W. Bauman, G. Zamora, M. S. Pattichis, and P. Soliz, "Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images," Invest. Ophthalmol. Vis. Sci. **52**(8), 5862–5871 (2011).
16. V. Sundaresan, K. Ram, K. Selvaraj, N. Joshi, and M. Sivaprakasam, "Adaptive super-candidate based approach for detection and classification of drusen on retinal fundus images," Ophthalmic Medical Image Analysis Workshop (OMIAW), MICCAI, (2015).
17. C. Agurto, H. Yu, V. Murray, M. S. Pattichis, S. Barriga, and P. Soliz, "Detection of hard exudates and red lesions in the macula using a multiscale approach," IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), 13–16 (2012).
18. M. Niemeijer, M. D. Abràmoff, and B. van Ginneken, "Information fusion for diabetic retinopathy CAD in digital color fundus photographs," IEEE Trans. Med. Imaging **28**(5), 775–785 (2009).
19. R. Klein, M. D. Davis, Y. L. Magli, P. Segal, B. E. K. Klein, and L. Hubbard, "The Wisconsin age-related maculopathy grading system," Ophthalmology **98**(7), 1128–1134 (1991).
20. C. P. Wilkinson, F. L. Ferris 3rd, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, J. T. Verdaguer, and Global Diabetic Retinopathy Project Group, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales,"Ophthalmology **110**(9), 1677–1682 (2003).
21. L. Wu, P. Fernandez-Loaiza, J. Sauma, E. Hernandez-Bogantes, and M. Masis, "Classification of diabetic retinopathy and diabetic macular edema," World J. Diabetes **4**(6), 290–294 (2013).
22. C. I. Sánchez, M. Niemeijer, I. Isgum, A. V. Dumitrescu, M. S. A. Suttorp-Schulten, M. D. Abràmoff, and B. van Ginneken, "Contextual computer-aided detection: Improving bright lesion detection in retinal images and coronary calcification identification in CT scans," Med. Image. Anal. **16**(1), 50–62 (2012).
23. M. Niemeijer, B. van Ginneken, J. Staal, M. S. A. Suttorp-Schulten, and M. D. Abràmoff, "Automatic detection of red lesions in digital color fundus photographs," IEEE Trans. Med. Imaging **24**(5), 584–592 (2005).
24. T. Kauppi, V. Kalesnykiene, J. K. Kamarainen, L. Lensu, I. Sorri, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "DIARETDB0 evaluation database and methodology for diabetic retinopathy algorithms," Tech. rep. (2006).
25. T. Kauppi, V. Kalesnykiene, J. K. Kamarainen, L. Lensu, I. Sorri, R. A., V. R., H. Uusitalo, H. Kälviäinen, and J. Pietilä, "DIARETDB1 diabetic retinopathy database and evaluation protocol,"Tech. rep. (2007).
26. M.D. Abramoff and M. Niemeijer, "Automatic detection of the optic disc location in retinal images using optic disc location regression," EMBS 4432–4435 (2006).
27. A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," IEEE Trans. Med. Imaging **19**(3), 203–210 (2000).
28. A. Rocha, T. Carvalho, H. F. Jelinek, S. Goldenstein, and J. Wainer, "Points of interest and visual dictionaries for automatic retinal lesion detection," IEEE Trans. Biomed. Eng. **59**(8), 2244–2253 (2012).
29. M. Niemeijer, M. D. Abràmoff, and B. van Ginneken, "Fast detection of the optic disc and fovea in color fundus photographs," Med. Image. Anal. **13**(6), 859–870 (2009).
30. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **2**, 2169–2178 (2006).
31. F. Samuelson, and N. Petrick, "Comparing image detection algorithms using resampling," ISBI 1312-1315 (2006).
32. D. Sidibé, I. Sadek, and F. Mériaudeau, "Discrimination of retinal images containing bright lesions using sparse

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

coded features and svm," Comput. Biol. Med. **62**, 175–184 (2015).
33. R. Pires, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Advancing bag-of-visual-words representations for lesion classification in retinal images," PLoS One **9**(6), e96814 (2014).
34. M. J. J. P. van Grinsven, A. Chakravarty, J. Sivaswamy, T. Theelen, B. van Ginneken, and C. I. Sánchez, "A bag of words approach for discriminating between retinal images containing exudates or drusen," ISBI 1444–1447 (2013).
35. P. P. Srinivasan, L. A. Kim, P. S. Mettu, S. W. Cousins, GR. M. Comer, J. A. Izatt, and S. Farsiu, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," Biomed. Opt. Express **5**(10), 3568–3577 (2014).
36. Y.-Y. Liu, H. Ishikawa, M. Chen, G. Wollstein, J. S. Duker, J. G. Fujimoto, J. S. Schuman, and J. M. Rehg, "Computerized macular pathology diagnosis in spectral domain optical coherence tomography scans based on multiscale texture and shape features," IOVS **52**(11), 8316–8322 (2011).

1. Introduction

Diabetic retinopathy (DR) is the leading cause of blindness worldwide in the working population with an estimated number of affected patients of 93 million in 2010 [1–3]. Age-related macular degeneration (AMD) is another sight threatening disease and the most common cause of blindness in the elderly. The estimated number of patients affected by AMD was 170 million in 2014 [4]. Both diseases progress without any visual complaints in early stages, while leading to visual impairment and, ultimately, vision loss in advanced stages. Major risk factors for these diseases include diabetes and an increased age. With the rising prevalence of diabetes and the aging population, the incidence of DR and AMD is expected to increase rapidly in the near future [5, 6].

To improve early detection of these diseases, screening for both DR and AMD through retinal examination using color fundus (CF) images is recommended by national authorities [7, 8]. However, only screening for DR is currently routinely implemented for diabetic patients. The main limitation of widespread implementation of DR and AMD screening is the enormous screening population and the associated costs of acquiring and analyzing the large amount of images generated. Automatic software solutions have been proposed to allow for more cost-effective mass-screening, reducing the amount of specialized personnel required and making mass-screening feasible. Most of these automatic software solutions analyze CF images for presence of lesions which are associated with DR or AMD [9–17]. Lesions associated with DR include microaneurysms, hemorrhages, exudates and cotton wool spots, whereas for AMD, these include drusen. After fusing the information obtained from the individual lesion detections, a final decision is made and in case of a positive finding, the patient is referred to an ophthalmologist for a precise diagnosis and follow-up procedure [18].

To make a correct decision for patient referral, it is evidently important to identify which types of lesions are present. On CF images, drusen have a similar bright appearance as exudates and cotton wool spots. Whereas cotton wool spots can be more easily identified by their generally larger size and different color appearance, exudates and drusen have very similar characteristics and differentiation is difficult. Following the criteria of international grading schemes for DR and AMD, a patient referral decision is different when a patient presents with drusen or exudates [19–21]: A patient can have a few drusen present without the need for a direct referral to an ophthalmologist, whereas the presence of only a few exudates is an indication for more severe levels of DR and the patient should be referred immediately.

As the underlying cause and disease mechanisms for DR and AMD are different, exudates and drusen appear with different location patterns in the retina. Exudates are caused by leaking fatty deposits from blood vessels and appear in compact groups, whereas drusen are believed to be a result of a reduced capacity of the retina to cleanse waste products from the photoreceptors and can appear over the whole retina. In our previous work we have shown that including con-

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

textual information is beneficial for individual lesion detection [22]. We believe that including spatial information is an important factor to differentiate between drusen and exudates.

In this work, we therefore propose an automatic classification scheme that employs spatial information to both identify and discriminate between images containing drusen and exudates. The method makes use of information of individual lesions as detected by three separate systems focusing on the detection of red lesions, bright lesions in general, and drusen in particular. We introduce a concentric spatial pyramid approach with a twofold purpose: 1) to incorporate spatial information in the classification step and 2) to structurally transform lesion-level information into image-level information. Features encoding local lesion load are computed in an increasingly fine concentric spatial grid and used in combination with a random forest classifier to make the final three-class classification. We compare our system performance with that of two independent human graders.

2. Materials and Methods

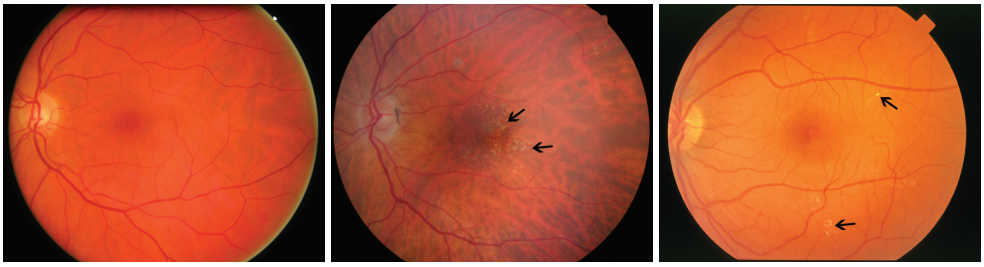
The automatic method to identify and differentiate between drusen and exudates consists of four steps. First, to standardize image resolution across different datasets, the field of view is automatically extracted and resized to 650 pixels in diameter. Next, three lesion detection and classification algorithms are applied to the images for the detection and classification of: (1) bright appearing lesions [14], i.e. hard exudates, cotton wools spots and drusen; (2) drusen [9]; and (3) red lesions [23], i.e. microaneurysms and hemorrhages. In the third step, results of these systems, consisting of the detected lesions and their associated posterior probability of being a true lesion, are combined. After removing lesions with a low posterior probability, the detected lesions are used as inputs to the spatial pyramid framework by creating histograms encoding the lesion load for each of the different type of detected lesions. Red lesion information is also included as the presence of red lesions is considered to be an indicator of DR, and therefore bright appearing lesions are more likely to be exudates. Finally, a multi-class classification is obtained by using a random forest classifier in a one-versus-all classification scheme using the spatial pyramid features. Using such a general framework allows to incorporate contextual information for differentiation between lesion types and allows to transform lesion-level information into image-level information.

2.1. Study Dataset

For this study, images were taken from several public datasets as well as from private datasets. In total, 130, 89, 397, 1200 and 569 images were taken from DiaretB0 [24], DiaretB1 [25], Stare [27], Messidor (Kindly provided by the Messidor program partners, see http://messidor.crihan.fr) and DR1/DR2 [28] datasets, respectively. After adding 488 and 1777 images, which were consecutively selected from the European Genetic Database (EUGENDA) (http://www.eugenda.org) and from the diabetic screening database of KSYOS TeleMedical Center (http://www.ksyos.org), we obtained a total of 4650 images. Images from the DiaretB0 and DiaretB1 dataset were taken at 50^o field of view with a resolution of 1500x1152 pixels, while the type of camera is not reported. Images from the Stare dataset were taken with a TopCon TRV-50 fundus camera at 35^o field of view and subsequently digitized at 605x700 pixels in resolution. Images from the Messidor dataset were captured with a Topcon TRC NW6 nonmydriatic retinograph with a 45^o field of view and had a resolution of 1440x960, 2240x1488 or 2304x1536 pixels. Images from the DR1/DR2 dataset were captured using a Topcon TRC50X mydriatic camera with a 45^o field of view and 640x480 pixel resolution. Images from the EUGENDA dataset were taken with either a Topcon TRC501X at 50^o field of view or with a Canon CR-DGi at 45^o field of view. Resolution of the EUGENDA images varied between 1360x1024

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

and 3504x2336 pixels. Images from the DR screening dataset had a resolution varying between 1024x768 and 4992x3328 pixels. Camera type and field of view were not reported for this set.



(a) (b) (c)

Fig. 1. Examples of color fundus images. (a): Control case, (b): image showing the presence of drusen and (c): image showing the presence of exudates.

Images which were not macula centered, had insufficient quality, or contained other bright appearing abnormalities such as myelinated nerve fibers or abnormalities not related to DR or AMD were discarded from the study dataset. The number of discarded images for each of these criteria was 1002, 270 and 302, respectively. The remaining set of images was graded by a human expert, who had over ten years of experience, into one of following categories: control, containing only drusen, containing only exudates, or containing both drusen and exudates. All images containing either drusen (N=362) or exudates (N=186) as graded by the expert were included and complemented with randomly selected images graded as control (N=552), resulting in a final study dataset of 1100 images of 1100 eyes. Figure 1 shows example images of the three different classes. Gradings of the human expert were considered as the reference standard in this study. Table 1 shows the distribution of the images among the different data sources. Two independent human observers, referred to as Observer 1 and Observer 2, also graded this set and rated every image into one of three categories: control, containing drusen or containing exudates. Observer 1 and Observer 2 had four and six years of experience, respectively. In addition, Observer 1 also identified subtle cases within the study dataset, where his/her confidence level in discriminating the underlying disease was low. Image quality, image focus, lesion size and lesion number influenced the level of confidence of the observer.

2.2. Preprocessing

In a preprocessing step, all images were resized to have the same field of view diameter of 650 pixels. No correction of illumination or shade correction was applied to the images. The optic disk location and fovea location were automatically detected using a previously developed algorithm [29], and were used later in the spatial pyramid framework.

2.3. Automatic lesion identification

Three types of lesions were automatically detected using previously developed algorithms: (1) Bright appearing lesions, i.e. hard exudates, cotton wools spots and drusen [14]; (2) drusen [9]; and (3) red lesions, i.e. microaneurysms and hemorrhages [23]. The systems generally consisted of a candidate extraction step, followed by a candidate classification step. Features based on shape, color and intensity were extracted from each candidate and used in a supervised classification framework to classify each individual candidate as being a true lesion. Table 2 shows a brief overview of the different features used for lesion classification. Candidates which overlapped with the optic disc location were automatically discarded in these previous works

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

Table 1. Distribution of images across different data sources. Only macular centered images with sufficient grading quality showing presence of only drusen or exudates were included and complemented with random control images.

	DB0	DB1	Eugenda	KSYOS	Messidor	Stare	DR1/DR2	Total
Control	18	10	47	142	306	5	24	552
Drusen	2	1	163	92	68	21	15	362
Exudates	25	14	1	12	99	25	10	186
Total	45	25	211	246	473	51	49	1100

DB0: DiaretB0; DB1: DiaretB1.

using an automatic optic disc detection system [26,29]. More details on the individual system's implementations can be found following the provided references [9, 14, 23]. The outputs of these systems are the detected bright appearing lesions, drusen and red lesions, respectively, and their associated posterior probability of being a true lesion. Figure 2 shows examples of the individual system outputs.

Table 2. Features for classification of lesion candidates as used the works for bright appearing lesion detection, drusen detection and red lesion detection [9, 14, 23].

Feature type	**Criteria**
Shape	Area, perimeter, compactness, length and width of the lesion candidate.
Context	Average and standard deviation of vessel pixel probability at the lesion candidate border. Distance to the closest lesion candidate. Number and average pixel probability of neighboring lesion candidates within a certain radius.
Intensity	Features measuring the difference in intensity in the RGB color space of the lesion candidate as compared to its surrounding area. Mean and standard deviation of Gaussian filter bank outputs.
Color	Average and standard deviation inside and outside the lesion candidate using the planes of the Luv and HSI color space.
Misc.	Average, standard deviation, maximum and median pixel probability inside the lesion candidate.

HSI: hue-saturation-intensity; Luv: luminescence-saturation-hue angle color space adopted by the International Commission on Illumination (CIE); RGB: red-green-blue.

2.4. Feature descriptor

Using the three lesion type probabilities, an estimate of the lesion load was made. This was done by constructing a weighted histogram of the detected lesions taking the size of the lesion into account. The size of the detected lesions was included as exudates are in general smaller in size compared to drusen. Given the probability P_i of the detected lesion i, the value h_n of the histogram bin n is defined as:

$$h_n = \sum_{i \in L_n} P_i \tag{1}$$

where L_n is the group of lesions whose size is n pixels, calculated as the smallest enclosing diameter of the lesion. The number of bins was set to 25 as most lesions have sizes smaller than

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

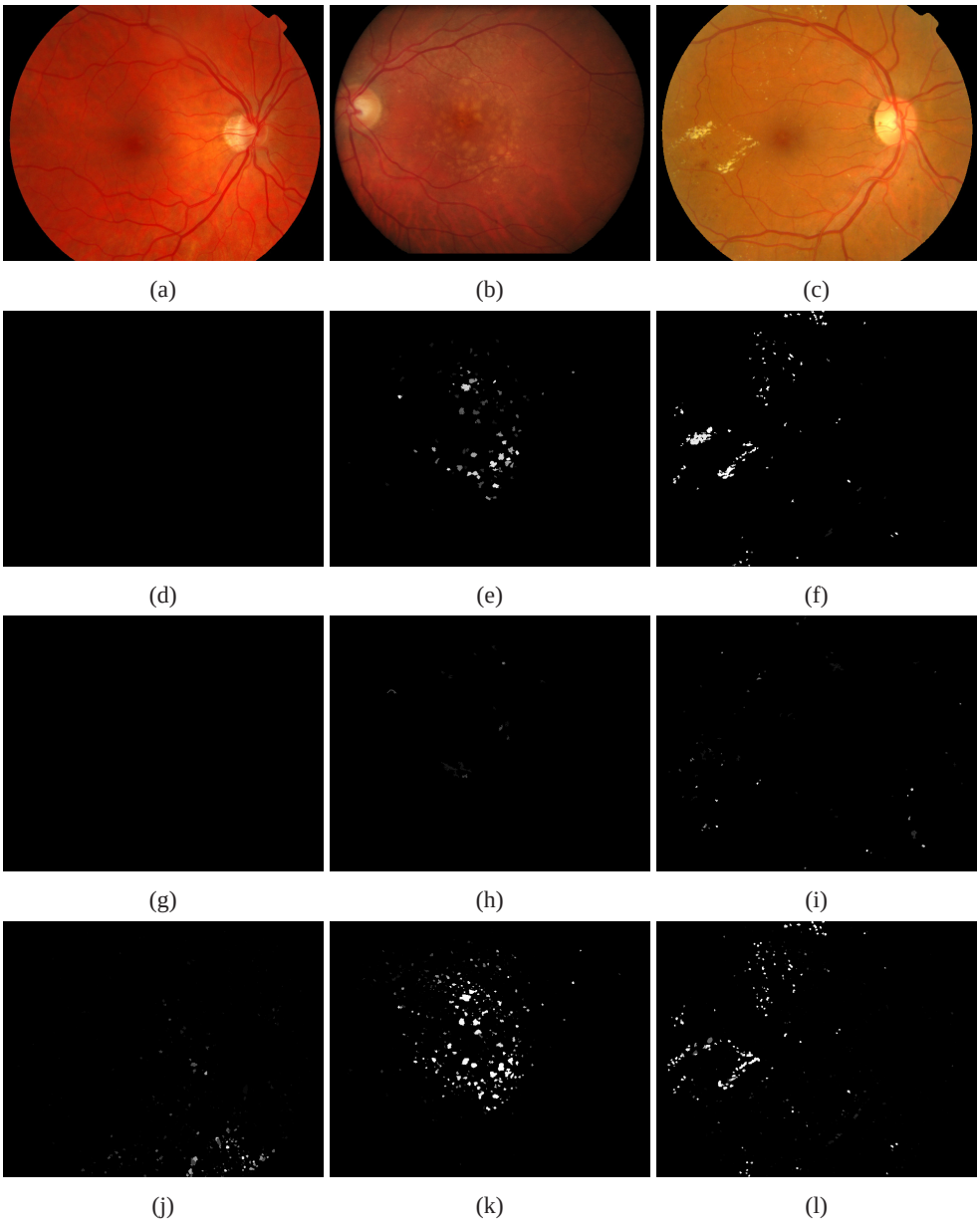


Fig. 2. Examples of images with the outputs of the individual automatic lesion detection systems. Top row: (a): Control case, (b): image showing the presence of drusen and (c): image showing the presence of exudates. Second row: output of the bright lesion detection system [14]. Third row: output of the red lesion detection system [23]. Bottom row: output of the drusen detection system [9]. In each of the lesion detection output maps, a brighter color indicates a higher likelihood of being a true lesion.

25 pixels. The last bin also takes into account lesions with size larger than 25 pixels. To remove false positive detections by the individual lesion identification systems, all bright appearing lesions and drusen with P_i smaller than 0.5 and all red lesions with P_i smaller than 0.3 were

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

neglected during construction of the weighted histograms. The values for these thresholds were determined in pilot experiments with each of the individual lesion detection systems.

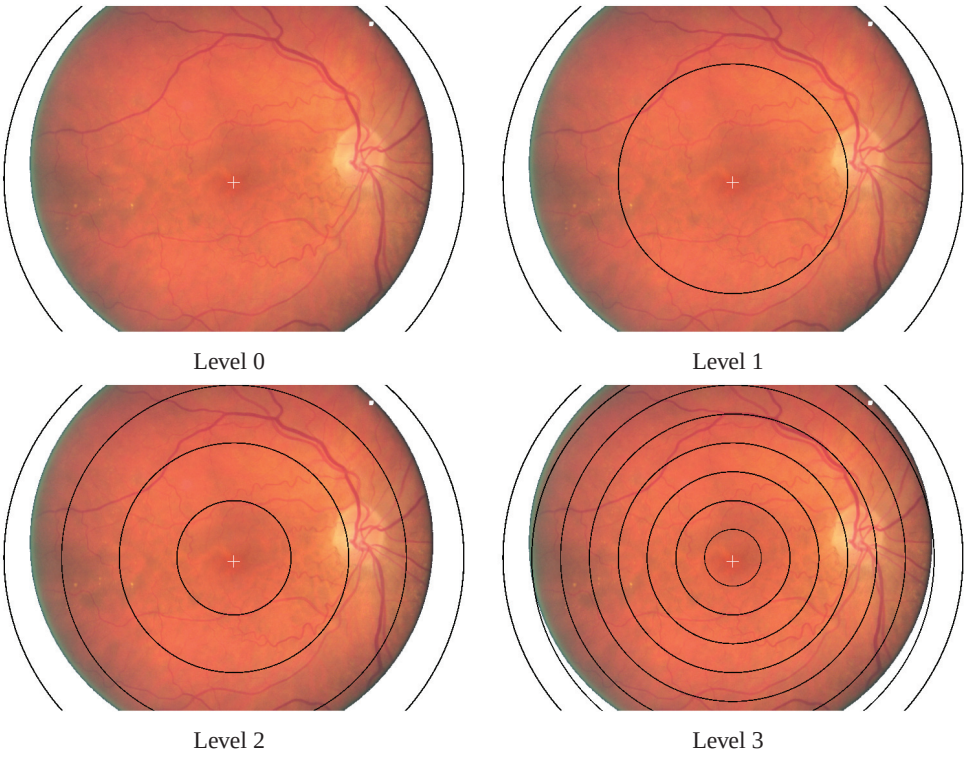


Fig. 3. Division of the image into concentric regions at different spatial pyramid levels. At each level, the number of concentric regions is doubled. Each region is centered on the fovea (depicted with the white cross) and the outer region has a radius of two times the distance between the fovea and optic disk.

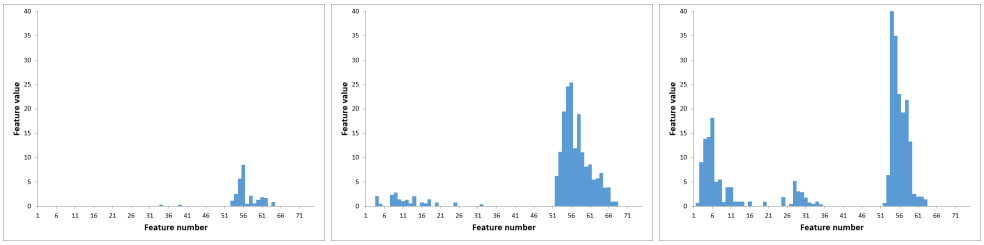
2.5. Spatial pyramids

The image was divided into increasingly coarser concentric circular regions [30]. With increasing coarseness, i.e. level, the number of concentric regions was doubled. Each circular region was centered on the automatically detected fovea location. At the 0^{th} level, the radius of the circular region was equal to twice the distance between the automatically detected fovea and the optic disc center. This radius was chosen in such a way that, for most images, the complete field of view of the image fell inside the 0^{th} level region. At each increasing level, the number of regions was doubled with equally spaced radii. See Fig. 3 for an example of the regions at different levels.

For each region, three lesion encoding weighted histograms, Equation 1, were computed including only lesions whose center of gravity fell inside the region. Histograms for each of the lesion types were concatenated, resulting in a feature vector of 75 features per region. For higher level regions, these regions can be seen as hollow circles, i.e. a lesion was included only once in a weighted histogram. Finally, one image encoding feature vector was composed for each image following either a single-scale approach or a multi-scale approach. In the single-scale approach, all feature vectors of the different regions at the level of evaluation were concatenated. In the

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

multi-scale approach, all feature vectors of regions at the current level and the feature vectors of the regions at the previous, lower levels, were concatenated. Figure 4 shows a graphical representation of the image feature vector computed at spatial pyramid level 0.



(a) (b) (c)

Fig. 4. Graphical representation of the image feature vector at the 0^{th} level spatial pyramid. (a): Feature vector of the image shown in Fig. 2(a), (b): feature vector of the image shown in Fig. 2(b) and (c): feature vector of the image shown in Fig. 2(c).

2.6. Multi-class differentiation of controls, drusen and exudates

Classification of the image into either control, containing drusen, or containing exudates was done using a one-versus-all classification scheme using the constructed image feature vector and a random forest classifier. In each of the one-versus-all classifications, 100 decision trees with a tree depth of 18 were used for the random forest classifier. All experiments were performed in a 10-fold cross-validation scheme. The total dataset was split up into 10 subsets. For each of the 10 folds, the random forest classifiers in the one-versus-all classification scheme were trained using 9 subsets and the left out subset was classified. Rotating this scheme 10 times resulted in classification of all subsets, without training and testing on the same data. All classified subsets were then pooled back together and performance was calculated by assigning the label of the class with the highest assigned probability. Receiver Operating Characteristics (ROC) analysis was performed using the assigned soft-probabilities for each of the individual classes in a one-versus-all classification scheme. Area (Az) under the ROC curve and 95% confidence intervals were calculated using bootstrap analysis [31].

3. Results

3.1. Observer grading

Two independent observers independently graded all images into one of the three categories: control, drusen or exudates. Results of the human observers are reported in a contingency table, shown in Table 3.

Table 3. Contingency table showing the results for the human observer gradings.

Reference	Observer 1			Observer 2		
	Control	Drusen	Exudates	Control	Drusen	Exudates
Control	479	55	18	404	138	10
Drusen	33	302	27	13	340	9
Exudates	1	2	183	0	11	175

Observer 1 had an overall accuracy of 0.876 and kappa agreement of 0.802 with 95% confidence interval (CI) of [0.771;0.833], whereas observer 2 had an overall accuracy of 0.835

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

and kappa agreement of 0.740 [0.706;0.775]. Eyes that were graded as control by the reference and as having drusen by the observers contributed mostly to the errors made by the observers. Individual sensitivity/specificity pairs for differentiating controls, drusen and exudates were computed in a one-versus-all fashion and are reported in Table 6.

3.2. Automatic differentiation between controls, drusen and exudates

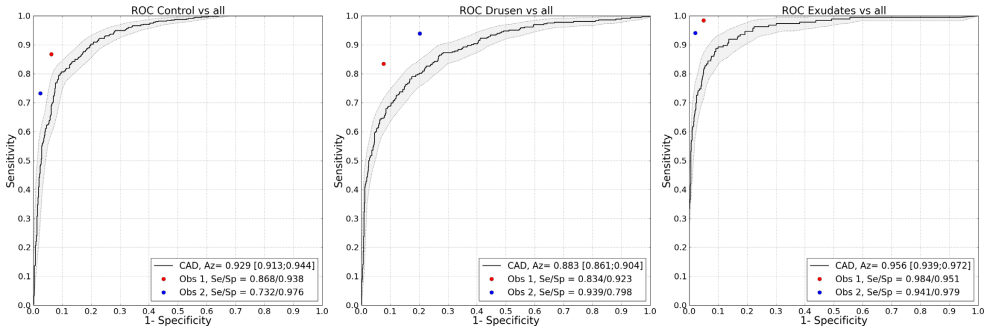
System performance was measured by applying two spatial pyramid schemes: one using single-scale analysis and one using multi-scale analysis. Table 4 shows the results for the single- and multi-scale spatial pyramid approaches. Kappa agreement and overall accuracy is reported.

Table 4. Classification performance of the automatic system using single- and multi-scale spatial pyramid approaches.

	Single-scale		**Multi-scale**	
Pyramid level	Kappa [95% CI]	Accuracy	Kappa [95% CI]	Accuracy
0	0.666 [0.626;0.705]	0.801	0.666 [0.626;0.705]	0.801
1	0.665 [0.625;0.705]	0.800	0.664 [0.624;0.703]	0.799
2	0.635 [0.594;0.676]	0.785	0.663 [0.623;0.703]	0.799
3	0.608 [0.566;0.651]	0.772	0.663 [0.623;0.703]	0.799

CI: Confidence interval.

Figure 5 shows the ROC curves for each of the one-versus-all classifications. The operating points of both human observers are added to each graph. The CAD system achieved an area Az under the ROC curve of 0.929, 0.883 and 0.956 for the detection of controls, images containing drusen and images containing exudates, respectively. Table 5 shows the contingency table for the automatic system for differentiation between controls, drusen and exudates at the 0^{th} pyramid level. Note that the results for the 0^{th} level was the same for the single- and multi-scale approaches. The multi-class approach achieved a kappa agreement of 0.666 [0.626;0.705] and accuracy of 0.801. Table 6 shows the sensitivity and specificity of the automatic system for identifying controls, drusen or exudates. The 0^{th} level spatial pyramid approach achieved sensitivity/specificity pairs of 0.904/0.781, 0.699/0.896 and 0.694/0.976 for controls, drusen and exudates, respectively.



(a) (b) (c)

Fig. 5. Receiver operating characteristics curves for each of the one-vs-all classifications. Area (Az) under the curve and 95% confidence intervals are reported in the legend. (a): Control versus all, (b): drusen versus all and (c): exudates versus all.

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

Table 5. Contingency table showing the results for the multi-class automatic system differentiation between controls, drusen and exudates using features of the 0^{th} pyramid level.

Reference	Automatic system		
	Control	Drusen	Exudates
Control	499	42	11
Drusen	98	253	11
Exudates	22	35	129

Table 6. Performance of the observers and the automatic spatial pyramid approach using features of the 0^{th} pyramid level.

Grader	Accuracy	Kappa [95% CI]	Control	Drusen	Exudates
Obs 1	0.876	0.802 [0.771;0.833]	0.868/0.938	0.834/0.923	0.984/0.951
Obs 2	0.835	0.740 [0.706;0.775]	0.732/0.976	0.939/0.798	0.941/0.979
System	0.801	0.666 [0.626;0.705]	0.904/0.781	0.699/0.896	0.694/0.976

Obs 1: Observer 1, Obs 2: Observer 2, System: automatic system.
Control, Drusen, Exudutes: Sensitivity/specificity pair for identification of the corresponding class using one-vs-all approach.
CI: Confidence interval.

3.3. Evaluation of influence of image subtlety

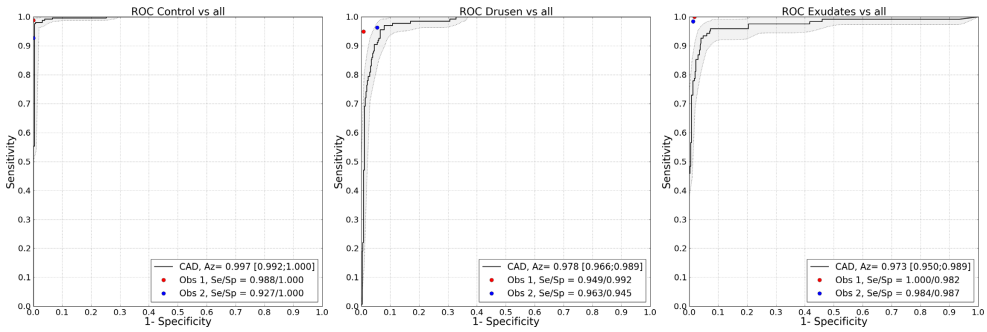
Observer 1 indicated for each image whether he/she was confident about his/her grading. Image gradings of which observer 1 was less confident were marked as subtle cases (N=582), whereas others were marked as clear cases (N=518). Table 7 shows the performance of the observers and the spatial pyramid method on the subset of subtle and clear cases. Total number of subtle cases was 292, 226 and 64 for normal, drusen and exudates, respectively. For clear cases, these numbers were 260, 136 and 122 for normal, drusen and exudates, respectively. Figure 6 shows the ROC curves obtained using only the clear cases and the 0^{th} level spatial pyramid. The CAD system achieved Az values of 0.997, 0.973 and 0.978 for the detection of controls, images containing drusen and images containing exudates, respectively.

Table 7. Performance of the observers and the proposed spatial pyramid approach on the subset of subtle (N=582) and clear (N=518) cases.

Grader	Accuracy	Kappa [95% CI]	Control	Drusen	Exudates
Subtle cases					
Obs 1	0.784	0.645 [0.590;0.700]	0.760/0.883	0.765/0.848	0.953/0.927
Obs 2	0.734	0.564 [0.505;0.623]	0.558/0.955	0.925/0.640	0.859/0.973
System	0.687	0.444 [0.377;0.511]	0.842/0.590	0.549/0.865	0.469/0.971
Clear cases					
Obs 1	0.981	0.969 [0.950;0.988]	0.988/1.000	0.949/0.992	1.000/0.982
Obs 2	0.950	0.921 [0.891;0.950]	0.927/1.000	0.963/0.945	0.984/0.987
System	0.929	0.886 [0.850;0.921]	0.973/0.996	0.949/0.924	0.811/0.982

Obs 1: Observer 1, Obs 2: Observer 2, System: automatic system.
Control, Drusen, Exudutes: Sensitivity/specificity pair for identifying the corresponding class.
CI: Confidence interval.

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA



(a) (b) (c)

Fig. 6. Receiver operating characteristics curves for each of the one-vs-all classifications using the set consisting of clear cases (N=518). Area (Az) under the curve and 95% confidence intervals are reported in the legend. (a): Control versus all, (b): drusen versus all and (c): exudates versus all.

4. Discussion

Correctly identifying the type of bright lesions is important for correct disease diagnosis in a screening setup as presence of drusen or exudates result in different referral criteria and might alter automatic referral decision of a patient. An automatic method using information of individual lesion detections in a global spatial pyramid framework to identify and differentiate between drusen and exudates was presented and evaluated in a heterogeneous dataset consisting of images from multiple sources and cameras. Using spatial information, the automatic method approaches the level of performance of human observers. The automatic system achieved an overall accuracy of 0.80 and kappa agreement of 0.67 with the reference. Observers achieved slightly higher performances with accuracies of 0.88 and 0.84 and kappa agreements of 0.80 and 0.74, respectively. Higher level spatial pyramids, in which the image was divided into more concentric sub-regions, did not lead to an increase in performance. Using multi-scale analysis is beneficial over single-scale analysis when using higher levels of spatial pyramids, but does not increase overall performance.

In our proposed framework an image level decision was generated by combining individual lesion detections in the whole image. Other methods have used a rule-based approach where the image level score was determined by the maximum lesion probability of the target class [9, 14]. To investigate the added value of our approach over such maximum rule based approach, we have applied the bright lesion detection system [14] and the drusen detection system [9] in our dataset. The bright lesion detection system achieved Az values of 0.906, 0.702 and 0.804 for the identification of controls, cases with drusen and cases with exudates, respectively. The drusen detection system achieved an Az value of 0.711 for the identification of cases with drusen. Our proposed method achieved substantially higher performance, with Az values of 0.929, 0.883 and 0.956 for controls, drusen and exudates, respectively. By not using only the maximum lesion detection response for the image score, our method is more robust to outliers and confounding lesions present in the data (i.e. drusen are confounding lesions when identifying cases with exudates and vice versa).

From Table 3 and Table 6, it can be observed that both observers achieved good performance for the identification of each of the individual classes. Control images which were graded as having drusen contributed most to the errors made by both observers. This can also be seen in the sensitivity for identifying control cases and specificity for drusen cases, which is slightly

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

lower than the ones for the other classes for both observers, see Table 6. A possible explanation for this is that drusen can be very subtle and small reflections in the retina can be misinterpreted as drusen. For exudates, this poses less of a problem as exudates in general have a higher contrast with the background. The proposed spatial pyramid approach achieved a sensitivity of 0.904 for identification of normal cases with a slightly lower specificity than observers. Lower sensitivities of the automatic system compared to observer for the classes drusen and exudates can be observed in Table 6.

One reason for the lower sensitivities of the automatic detection of images with drusen or exudates is the fact that the three automatic systems that detected various lesions are not perfect. The outputs of these systems are used in the spatial pyramid framework. Lesions which are not detected by these systems will not be incorporated in the spatial framework and hamper thus the ability of the automatic system to make a correct decision. This is reflected in the larger number of images assigned to the control class, while being a drusen or exudates case, see Table 5, and the lower sensitivities for identification of drusen and exudates cases, see Table 6.

Lowering the threshold for the lesions to be included in the framework increases the chance of subtle lesions to be included in the spatial pyramid framework, but also contributes to a larger number of false positive detections. The thresholds were determined using pilot experiments with each of the individual lesion detection systems. Additional analysis of the influence of these system hyper-parameters on the final performance of our proposed system showed similar performance. Table 8 shows the accuracy values obtained by the 0^{th} pyramid level CAD system. A change in threshold values led to only minor changes on the overall CAD system's accuracy. However, these thresholds might not be optimal for the identification of the individual classes, such as drusen or exudates.

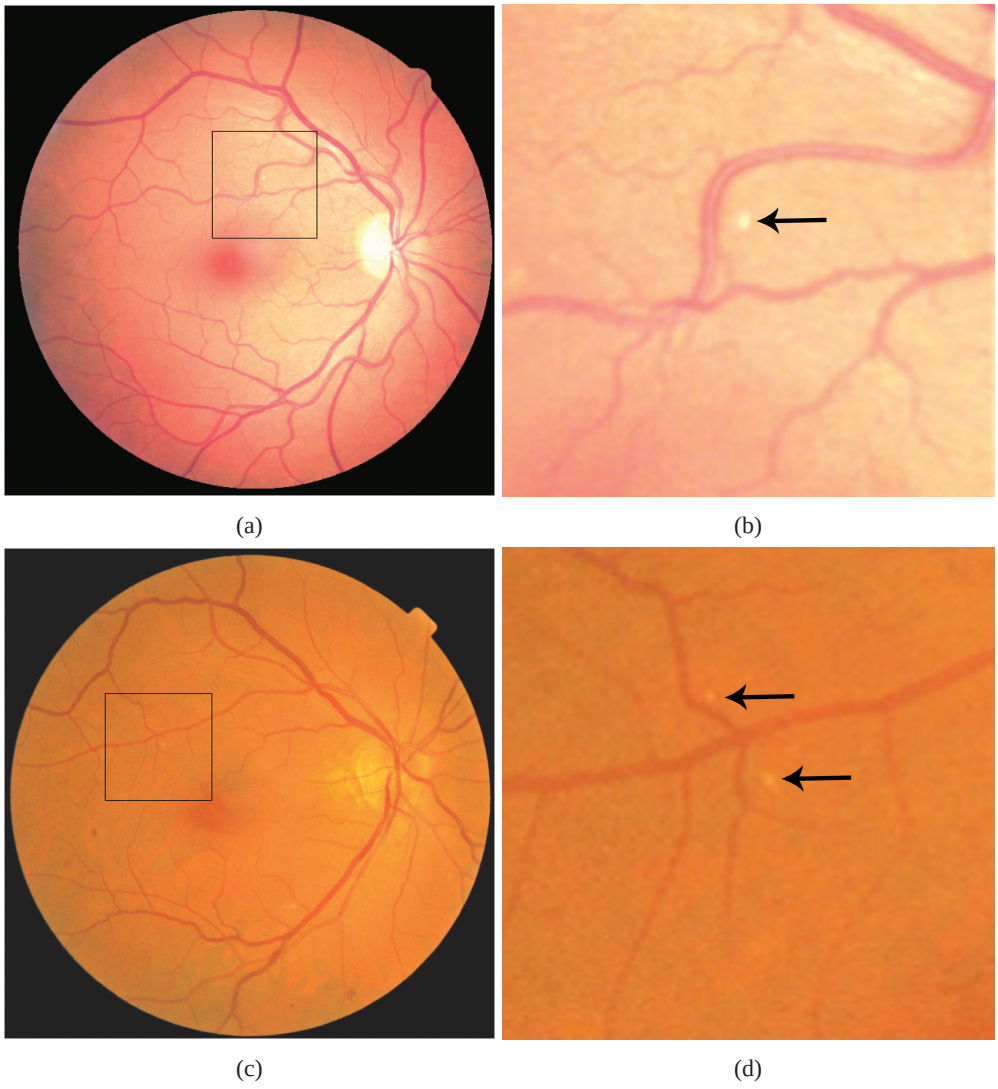
Table 8. Accuracy values obtained using the 0^{th} pyramid level CAD system using different threshold settings for discarding bright appearing lesion, drusen and red lesion information. Threshold 1: threshold for discarding bright appearing lesions and drusen; Threshold 2: threshold for discarding red lesions.

Accuracy		Threshold 2				
		0.0	0.1	0.2	0.3	0.4
Threshold 1	0.0	0.795	0.804	0.800	0.803	0.795
	0.1	0.792	0.796	0.789	0.793	0.799
	0.2	0.791	0.797	0.798	0.795	0.793
	0.3	0.799	0.809	0.799	0.798	0.797
	0.4	0.799	0.803	0.803	0.801	0.795
	0.5	0.800	0.812	0.810	0.801	0.803
	0.6	0.804	0.812	0.800	0.800	0.801

Another reason for the fact that the automatic system is performing worse than human experts may be the large amount of subtle cases in the dataset. Many images only contain a single or only a few drusen or exudates. Figure 7 shows examples of subtle cases included in the dataset. Images presenting with only a single subtle lesion are more likely to be missed by the automatic lesion detection systems. Furthermore, based on a single lesion, it is hard or even impossible for the automatic system to make a differentiation between drusen or exudates using spatial information. Table 5 shows the number of images correctly and incorrectly classified by the automatic system. To analyze the influence of subtlety of the data on the system performance, we evaluated performance on the subset of images marked as clear cases by Observer 1. System performance and also observer performance is much higher in this subset compared to the subtle subset, Table 7, reaching an accuracy of 0.93 and kappa agreement of 0.89. The ability

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

to detect and differentiate normal, drusen and exudates cases appears to be heavily influence by the subtlety of the images under consideration.



(a) (b)

(c) (d)

Fig. 7. Examples of images showing subtle bright lesions. (a): image showing the presence of drusen with (b) showing a close up view of the region as indicated by the black box in (a). (c): image showing the presence of exudates with (d) showing a close up view of the region as indicated by the black box in (b). Both images (a) and (c) were graded by the automatic system as control.

In our study, we have excluded images if they contained myelinated nerve fibers or abnormalities which are not related to AMD or DR. However, in a screening setup, such images can be encountered. Therefore, we have investigated how our system would respond in such a case. Figure 8 shows an example of an image containing myelinated nerve fibers. The 0^{th} level spatial pyramid classified this image as containing exudates. Although this classification might not be correct, the system has identified that the image does not belong to the control class.

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

Although identification of this image as not being a control is important in a screening setting, an automated system which identifies multiple eye abnormalities would lead to a more accurate patient referral and treatment decision.

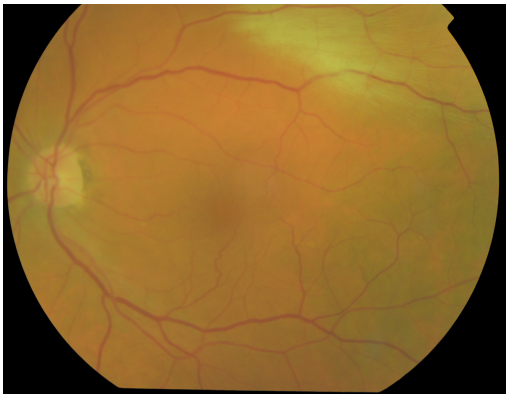


Fig. 8. Example of an image with myelinated nerve fibers.

Table 6 and Table 7 show that the incorporation of spatial information through spatial pyramids did not improve the final performance. The value of using higher spatial pyramids might not be well expressed for this task, but could potentially increase performance for other related classification tasks. In retinal screening, not only the type of lesions, but also the location of lesions is important to make a decision on whether a patient should be referred. Lesions close to the macular region represent a higher risk of disease progression for the patient and, therefore, influence the decision for referral. Higher level spatial pyramids could improve the performance as compared to the 0^{th} level as higher levels are capable of capturing this spatial information in the image. Another application where higher level spatial pyramids might prove to have added value is for disease severity grading. In the case of AMD severity grading, the level of severity is based on the location of the abnormalities with respect to the fovea. The 0^{th} level spatial pyramid is unable to encode this information as all lesions in the image are encoded in the same weighted histogram regardless their position. Higher level spatial pyramids are, however, able to capture this spatial information: the concentric circles of the higher level spatial pyramids are centered on the fovea and each ring encodes information at a fixed distance from the fovea. The property of encoding spatial information might have a higher impact on such classification tasks. We will further investigate the impact of higher level spatial pyramids for these tasks in future work.

We have used the ratings of a single expert as reference, mimicking the clinical protocol where a single observer grades the images. However, as the availability of multiple ratings for each image allowed for construction of a potentially more robust reference, we also examined the use of a consensus reference, constructed by using majority voting of the three observers. Training and evaluating the system using this consensus reference resulted in a similar performance with an accuracy of 0.793. One-versus-all sensitivity/specificity values of 0.883/0.817, 0.718/0.868 and 0.709/0.971 were achieved for the identification of controls, drusen and exudates, respectively. Using a consensus of the available graders to construct the reference omits us, however, to directly compare the results with independent human observers.

The choice of a suitable classifier is difficult to make in advance as the classifier performance depends on the dataset and the classification task at hand. We have also compared other classifiers such as a support vector machine and a k-nearest neighbor classifier, obtaining an accuracy of 0.736 and 0.696 for the 0^{th} level spatial pyramid, respectively. The random forest classifier

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

obtained higher performance with a moderate computational cost of 12 seconds for training and classification, computed on an Intel Xeon CPU with 2.4 GHz memory.

In literature, other methods which have focused on the identification of either drusen or exudates have been described [10–13, 15–17]. In these works, the focus was put on discriminating control cases from abnormal cases, i.e. cases containing either abnormalities related to AMD or abnormalities related to DR. The datasets for this task consisted of control cases and cases with either exudates or drusen, but not simultaneously. Features to differentiate between exudates and drusen need to be more robust as these abnormalities might have similar appearance on color fundus images, making differentiation more difficult. This difficulty is also reflected in Fig. 6 which shows that the CAD system achieves higher performance for the identification of controls with an Az value of 0.997, as compared to the identification of drusen (Az=0.978) or exudates (Az=0.973).

Previous works also described methods to discriminate between images containing bright lesions. A method, making use of sparse coded features and a support vector machine achieved near perfect discrimination results [32]. However, it should be noted that a fair comparison between methods is hard to make since results are based on different datasets. The composition of the datasets has a large influence on algorithm performance, as is evident from the difference in performance between subtle and clear cases in our experiments (see Table 7). In our study, images showing presence of single drusen or exudates were included. Based on a single lesion, it can be difficult or even impossible for an algorithm to make a differentiation. Reference criteria for data inclusion in the previous study [32] were not mentioned by the authors. Furthermore, no human observer was used in the previous study, which would allow to compare system performance to human level performance, giving a measure of overall performance. Including human observers is therefore important to draw conclusions on algorithm performance.

Other studies have proposed frameworks based on bag-of-visual-words (BoVW) representation which was used to detect bright lesions and red lesions [28, 33, 34]. In these approaches, a visual vocabulary is constructed using features extracted from the color fundus images. These methods achieve good performance on specific datasets. An evaluation of these methods on a diverse dataset consisting of image from multiple centers and cameras has not been performed. Furthermore, these methods have many system parameters which need to be tuned in order to obtain good results. These parameters include the number of visual words for image description and the manual crafted features used to map local image patches to one of the visual words. Furthermore, interpretation of the system is difficult as the visual vocabulary cannot be mapped directly to individual lesions or image characteristics.

In this work, we have combined features extracted from solely color fundus images. Adding information from different imaging modalities could, however, increase the CAD performance. For example, pathologies related to AMD and DR manifest at different retinal depths and affect different layers of the retina. As Optical Coherence Tomography (OCT) provides valuable depth information, features derived from OCT could potentially be beneficial for disease differentiation. Previous works employing OCT imaging data have shown good performance for the automatic differentiation of retinal diseases [35, 36]. A method incorporating histogram of oriented gradient features and a support vector machine correctly identified 100% of the cases with AMD, 100% of the cases with diabetic macular edema, and 86.7% of the normal subjects in a small study dataset of 15 subjects for each class [35]. Another work focused on the identification of different macular pathologies: macular hole, macular edema and AMD [36]. This method uses multiscale texture features and shape features in combination with support vector machines to identify control cases and each of the pathologies separately. The method obtained Az values ranging between 0.941 and 0.978 in a dataset comprising of 131 scans from 37 subjects. Extending our proposed framework by adding features derived from OCT data

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

could therefore potentially increase the classification performance. However, in current screening settings, only color fundus imaging is performed, which limits the inclusion of OCT derived features in our proposed framework.

To conclude, we presented an automatic system for the identification and differentiation of images containing drusen or exudates using a spatial pyramid framework. We provide a general framework to generate an image level decision based on individual lesion detections and we show that the automatic system approaches human level performance, although results do not improve when using higher level spatial pyramids. System performance is limited by the subtlety of the images in the dataset and leaves room for improvement. Improvements could include the addition of lesion level characteristics, such as color and shape in the spatial pyramid approach.

Acknowledgments

Supported by a ZonMw grant: "A cost-effective solution for the prevention of blindness using computer-aided diagnosis and fundus photography," with project number 11.631.0003.

#254550 Received 26 Nov 2015; revised 22 Jan 2016; accepted 23 Jan 2016; published 2 Feb 2016
(C) 2016 OSA

131.174.248.149 {ts '2018-03-19 05:55:08'}