Comparative study of retinal vessel segmentation methods on a new publicly available database

M. Niemeijer[ab], J. Staal[a], B. van Ginneken[a], M. Loog[a] and M.D. Abràmoff[abc]

[a]Image Sciences Institute, Univ. Medical Center Utrecht, Utrecht, The Netherlands
[b]Department of Ophthalmology, Vrije Univ. Medical Center, Amsterdam, The Netherlands
[c]Department of Ophthalmology and Visual Sciences, Univ. of Iowa, USA

ABSTRACT

In this work we compare the performance of a number of vessel segmentation algorithms on a newly constructed retinal vessel image database. Retinal vessel segmentation is important for the detection of numerous eye diseases and plays an important role in automatic retinal disease screening systems. A large number of methods for retinal vessel segmentation have been published, yet an evaluation of these methods on a common database of screening images has not been performed. To compare the performance of retinal vessel segmentation methods we have constructed a large database of retinal images. The database contains forty images in which the vessel trees have been manually segmented. For twenty of those forty images a second independent manual segmentation is available. This allows for a comparison between the performance of automatic methods and the performance of a human observer. The database is available to the research community. Interested researchers are encouraged to upload their segmentation results to our website (`http://www.isi.uu.nl/Research/Databases`). The performance of five different algorithms has been compared. Four of these methods have been implemented as described in the literature. The fifth pixel classification based method was developed specifically for the segmentation of retinal vessels and is the only supervised method. We define the segmentation accuracy with respect to our gold standard as the performance measure. Results show that the pixel classification method performs best, but the second observer still performs significantly better.

Keywords: vessel segmentation, retina, comparative study, image database

1. INTRODUCTION

Automatic segmentation of the vasculature in retinal images is important in the detection of a number of eye diseases. Some diseases, e.g. retinopathy of prematurity, affect the morphology of the vessel tree itself. In other cases, e.g. pathologies like microaneurysms, the performance of automatic detection methods may be improved if regions containing vasculature can be excluded from the analysis.[1,2] Another important application of automatic retinal vessel segmentation is in the registration of retinal images of the same patient taken at different times.[3] The registered images are useful in automatically monitoring the progression of certain diseases.[4] Finally, the position, size and shape of the vasculature provides information which can be used to locate the optic disk and the fovea.[5] Therefore the automatic vessel segmentation forms an essential component of any automated eye-disease screening system.

The research presented in this paper is part of a larger effort to develop an automated screening system for the detection of diabetic retinopathy in retinal images. A large number of methods for retinal vessel segmentation have been published, yet evaluation on a common database of screening images has not been performed. To compare the performance of vessel segmentation algorithms we have established a large database of retinal images in which the vessels have been segmented manually. This database is publicly available to the research community via a website. Researchers who use the data are encouraged to upload their segmentation results to this same website (for details see Appendix A).

Shared image repositories are already available for a variety of other medical image types. Examples are the JSRT chest radiography database[6] and the DDSM digital mammography database.[7] Hoover et al.[8] made their

The principal author can be reached via e-mail at: meindert@isi.uu.nl. For further information visit the Image Sciences Institute at http://www.isi.uu.nl/.

Medical Imaging 2004: Image Processing, edited by
J. Michael Fitzpatrick, Milan Sonka, Proceedings of SPIE Vol. 5370
(SPIE, Bellingham, WA, 2004) · 1605-7422/04/$15 · doi: 10.1117/12.535349

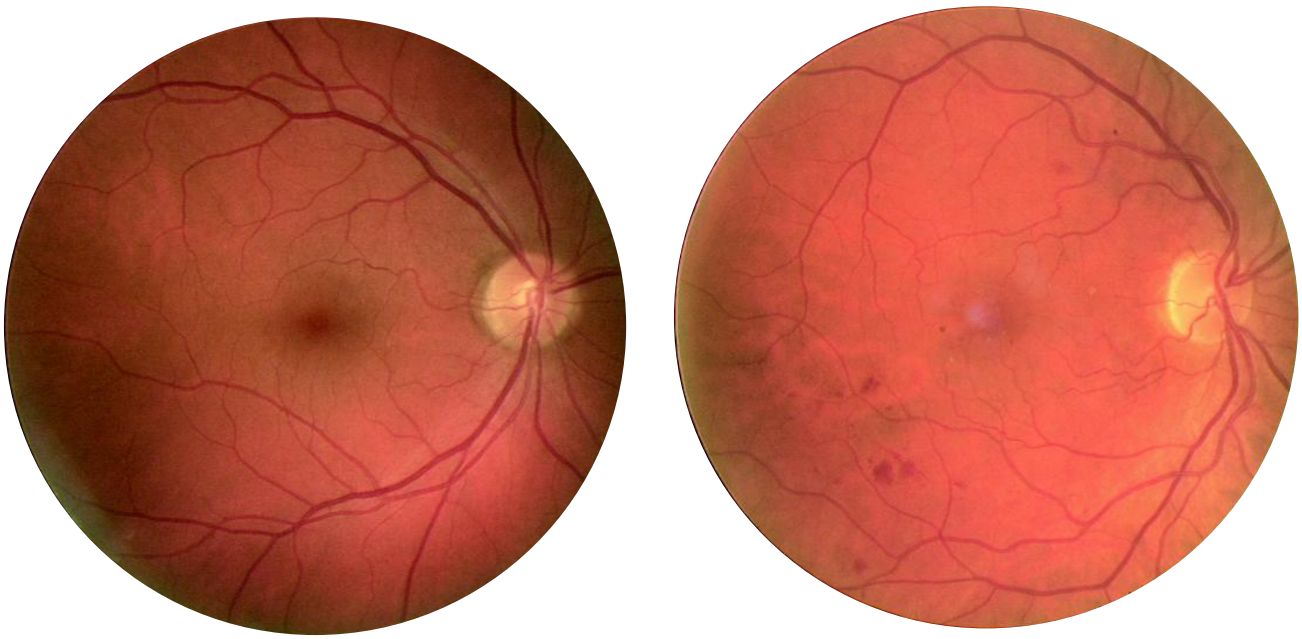

Figure 1. Left: A typical image from the database without pathology. Right: An example image from the database containing pathology (hemorrhages, microaneurysms). Note that only the FOV of both images is shown.

database of segmented retinal vessel images available on the web. However, there are several drawbacks to this set of images that motivated us to develop a new database as described in this work. First, the images in the Hoover database are not screening images. Because they are scanned from film the quality of the images is lower than if they would have been acquired directly using a digital fundus camera. Finally, the variability between the manual segmentation from the first and second observer is substantial.

In this work the performance of five vessel segmentation methods on a new publicly available image database is compared, four previously published methods and one new method.

2. MATERIALS

Our image database, to which we will refer to as the DRIVE (Digital Retinal Images for Vessel Extraction) database, consists of a total of 40 color fundus photographs. All images have been deidentified, they were stripped from all individually identifiable information and processed in such a way that this information cannot be reconstructed from the images. The photographs were obtained from a diabetic retinopathy screening program in The Netherlands. The screening population consisted of 453 subjects between 31 to 86 years of age. Each image has been JPEG compressed, which is common practice in screening programs. Of the 40 images in the database, 7 contain pathology, namely exudates, hemorrhages and pigment epithelium changes. See Figure 1 for an example of both a normal and a pathological image.

The images were acquired using a Canon CR5 non-mydriatic 3CCD camera with a 45 degree field of view (FOV). Each image is captured using 8 bits per color plane at 768 $\times$ 584 pixels. The FOV of each image is circular with a diameter of approximately 540 pixels.

The set of 40 images was divided into a test and a training set both containing 20 images. Three observers, the first and second author and a computer science student manually segmented a number of images. All observers were trained by an experienced ophthalmologist (the last author). The first observer segmented 14 images of the train set while the second observer segmented the other 6 images. The test set was segmented twice resulting in a set X and Y. Set X was segmented by both the first and second observer (13 and 7 images respectively) while set Y was completely segmented by the third observer. The performance of the vessel segmentation algorithms

is measured on the test set. In set X the observers marked 577,649 pixels as vessel and 3,960,494 as background (12.7% vessel). In set Y 556,532 pixels are marked as vessel and 3,981,611 as background (12.3% vessel).

3. METHODS

Five different (retinal) vessel segmentation methods were tested on the DRIVE database. The first four of these algorithms have been implemented as described in the literature, the fifth method was developed specifically for the purpose of retinal vessel segmentation. A short description of each of the five methods is given below. Figure 3 shows the output of each of the methods on one image from the database.

3.1. Matched Filter Approach

Chaudhuri et al.[9] note that the gray-level profiles of the cross-sections of retinal vessels have an intensity profile which can be approximated by a Gaussian. A two-dimensional matched filter approach is proposed to detect the vessels. Vessel segments at various orientations are detected by convolving the image with rotated versions of the matched filter kernel and retaining only the maximum response. At an angular resolution of 15°, a total of 12 convolutions are needed. The resulting image can be thresholded to produce a binary segmentation of the vasculature.

3.2. Scale-Space Analysis and Region Growing Approach

Martínez-Pérez et al.[10] use a combination of scale space analysis and region growing to segment the vasculature. Two features are used to characterize the blood vessels, the gradient magnitude of the image intensity $|\nabla I|$ and the ridge strength both at different scales. The ridge strength is determined by calculating the absolute largest eigenvalue $|\lambda_1|$ of the matrix of second order derivatives of the image intensity (the Hessian). To account for the difference in vessel width across the retina both these features are normalized by the scale s over the scale-space while retaining only the local maxima. The local maxima of the gradient magnitude γ and the local maxima of the ridge strength κ then become:

$$\gamma = \max_s \left[\frac{|\nabla I(s)|}{s} \right] \quad , \quad \kappa = \max_s \left[\frac{|\lambda_1(s)|}{s} \right] \tag{1}$$

The histograms of both features are used in the final region-growing step, in which the images pixels are divided into two classes, "vessel" and "non-vessel". This is accomplished by alternating the vessel and background region growing and lowering the feature thresholds after each iteration. This continues until no new pixels are added to either of the two classes.

3.3. Mathematical Morphology and Curvature Estimation Approach

The method proposed by Zana et al.[11] is a general vessel segmentation method based on the use of mathematical morphology. The algorithm itself consists of several morphological operations and can be divided into a number of steps:

1. Recognition of linear parts by computing the supremum of openings using a linear structuring element at different orientations.
2. Noise suppression by using a geodesic reconstruction of the supremum of openings into the original image.
3. Removal of different types of undesirable patterns by applying the Laplacian on the result of the previous step followed by a specially designed alternating filter.

The final result can be thresholded to produce a segmentation of the vasculature.

3.4. Verification-Based Local Thresholding Approach

Jiang et al.[12] propose an adaptive local thresholding framework based on a verification-based multithreshold probing scheme. Retinal vessels cannot be segmented by using a global threshold because of gradients in the background of the image. Instead Jiang et al. propose to probe the image with a number of thresholds. At each of the probed thresholds all binary objects in the thresholded image are extracted. By applying a classification procedure to the objects, only those that have vessel-like features will be retained. All the retained binary objects can be combined to form a binary vessel tree segmentation. The sensitivity of the method can be manipulated by changing the parameters of the classification procedure, i.e. making it less or more strict.

3.5. Pixel Classification Approach

As a comparison for the other systems which are all non-supervised (not trained with examples) we implemented a simple vessel segmentation method based on pixel classification. For each pixel in the image, a feature vector is constructed and a classifier is trained with these feature vectors. Features are extracted from the green plane of the retinal images only. In initial experiments we compared two types of features: the output of filters and the pixel values within a neighborhood. The filter output showed the best results.

Initial experiments were conducted in which three classifiers[13] have been compared, a kNN-classifier, a linear classifier and a quadratic classifier. Performance of the kNN-classifier was superior for all experiments, so this classifier has been selected. The final experiments were conducted using a system with a kNN classifier where $k = 30$. The filter features consisted of the Gaussian and its derivatives up to order 2 at scales s = 1, 2, 4, 8, 16 pixels, augmented with the original (greenplane) image. The total number of features is therefore 5×6+1=31. Each feature is normalized to zero mean and unit variance before classification.

Because of the use of a kNN classifier it was possible to perform a soft classification. This resulted in a probability map which for each pixel in the image indicates the probability that it is a vessel pixel. By thresholding the probability map a binary segmentation of the vasculature can be obtained.

4. EXPERIMENTS AND RESULTS

4.1. Experiments

A number of experiments were performed to compare the performance of each of the methods on the DRIVE database. In this work the performance is evaluated based on the maximum average accuracy (MAA) of the segmentation in comparison with the first-observer based gold standard. Before determining the MAA the result images produced by the algorithms were (when necessary) thresholded at an optimal threshold determined on the training set. Using these thresholded images the accuracy per image could be determined by taking the total number of correctly classified vessel and non-vessel pixels for each image and then dividing this sum by the total number of pixels in the FOV. To determine the MAA we take the average value of the accuracy for all images.

All methods except the second observer and the method by Martínez-Pérez et al. result in a soft classification. This means we can interpret the intensities of the pixels in the resulting image as posterior probabilities. By thresholding the result images at different levels sensitivity and specificity pairs are obtained that can be used to plot an ROC curve.[14] The method by Jiang et al. is a special case where the sensitivity/specificity of the method can be changed by manipulating the classification function (see Section 3.4). In this case the ROC-curve plots the fraction of vessel pixels actually classified as vessel pixels (true positive fraction) against the fraction of non-vessel pixels classified as vessel pixels (false positive fraction). A larger area under the curve (A_z) signifies a greater discriminatory ability of the segmentation method. When all pixels have been classified at random $A_z = 0.5$, while $A_z = 1.0$ would mean perfect discrimination.

In all experiments the test set as described in Section 2 was used for evaluating the accuracy of the methods. Our supervised pixel classification based method was trained on the training set.

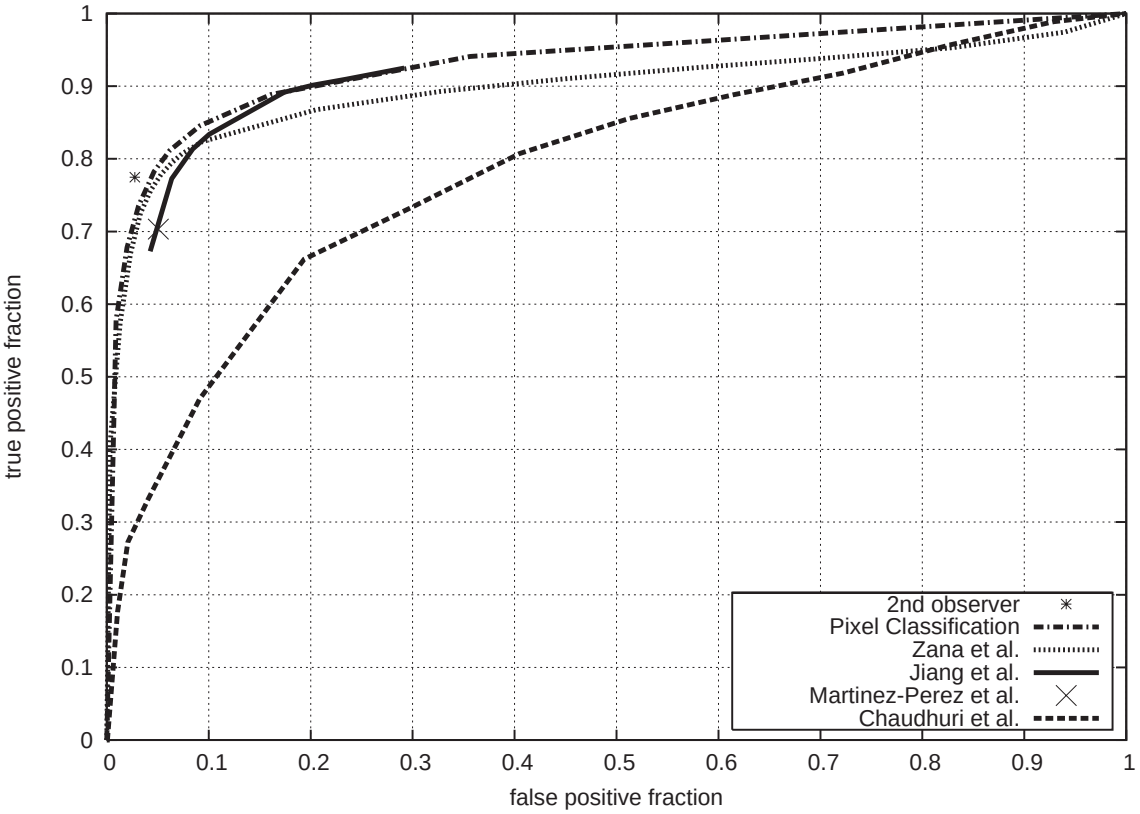

Figure 2. ROC curves of all tested methods.

4.2. Results

Figure 2 shows the ROC curves of the different methods. Both binary segmentation methods are shown as single points in the ROC-plot. Table 1 shows an overview of the results of the different methods. The accuracy given in Table 1 is the maximum average accuracy at a certain optimum threshold. This optimum threshold was determined on the training set. The Kappa value[15] in Table 1 is a measure for the agreement between two observers. In this case the agreement between the gold standard and each of the methods was measured. Finally the area under the ROC-curve is also given in the Table.

	Max. avg. Accuracy	Kappa	A_z
2^{nd} observer	0.9473 (0.0048)	0.7589	-
Pixel classification	0.9416 (0.0065)	0.7145	0.9294
Zana et al.	0.9377 (0.0077)	0.6971	0.8984
Jiang et al.	0.9212 (0.0076)	0.6399	0.9114
Martínez-Pérez et al.	0.9181 (0.0240)	0.6389	-
Chaudhuri et al.	0.8773 (0.0232)	0.3357	0.7878
All background	0.8727 (0.0123)	0	-

Table 1. Results of the different algorithms. The second column contains the maximum average accuracies with standard deviations. The third column contains Kappa (agreement) values. Finally the fourth column shows the area under the ROC curve. When we assign all pixels to the most likely class, in this case 'background' we obtain the accuracy as shown in the bottom row.

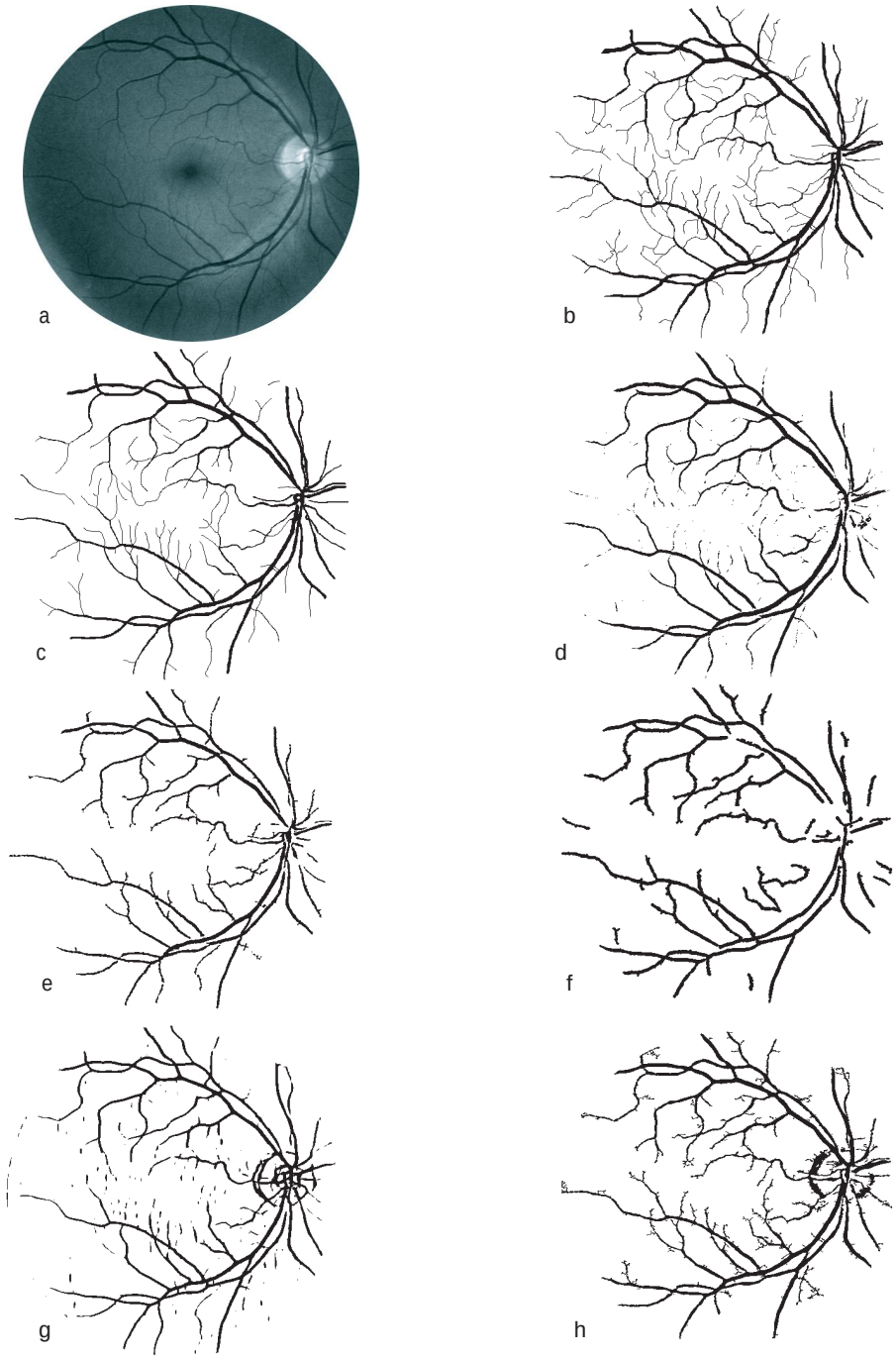

Figure 3. Segmentation results for the various methods, illustrated for one image from the test set. Where necessary, the result images were thresholded at the level where the average accuracy on the train set was maximum. (a) Original test image. (b) Gold standard. (c) Second manual segmentation. (d) Pixel classification method. (e) Zana et al. mathematical morphology and curve estimation method. (f) Jiang et al. verification-based local thresholding method. (g) Chaudhuri et al. matched filter method. (h) Martínez-Pérez et al. scale-space analysis and region growing approach.

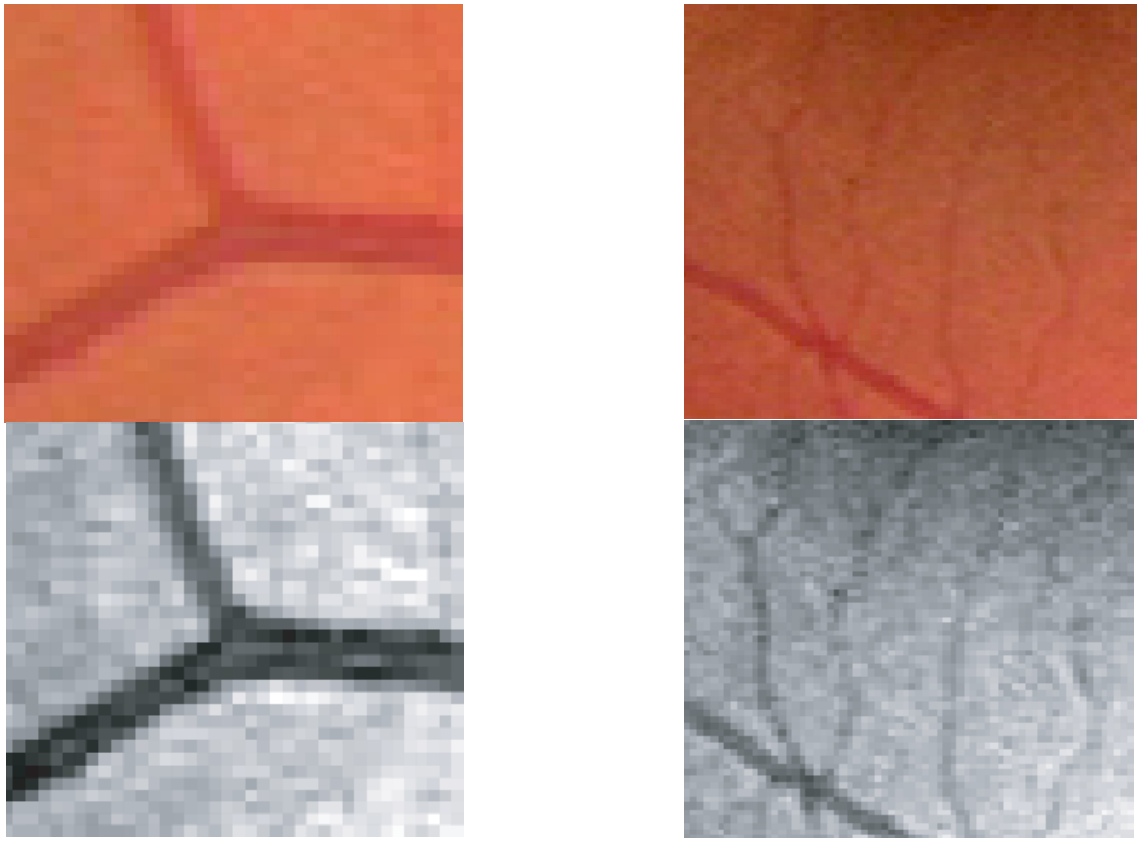

Figure 4. Left: Two images showing the same detail of a large vessel. The top image is the original, the bottom was window-levelled for maximum contrast. Note the blurry edges. Right: Detail of 4 small vessels moving from the lower part of the image to the upper part. Again both the original (top) and a window-levelled version (bottom) are shown.

5. DISCUSSION AND CONCLUSION

The difference between the first observer (gold standard) and the second observer shows that vessel segmentation is not an easy task. There are multiple factors complicating the manual segmentation of the images from the database. The smallest vessels are hard to see, especially if they are not wider than one pixel. JPEG compression artifacts further hamper the segmentation of those vessels. Another effect of the JPEG compression is that the location of the border of larger vessels becomes difficult to establish. Pixels that are included in the border of the vessel by the first observer might not be included by the second observer. Figure 4 shows examples of both effect. A final factor is the time involved in the manual segmentation, it takes on average 2 hours to manually segment one image using a simple painting tool. When segmentation times are this long, fatigue of the human observers can cause a loss in segmentation precision. In comparison with the most accurate automatic method (pixel classification) the second observer still is significantly more accurate, in a paired t-test $P < 0.001$. All other differences in accuracy between the methods are significant as well with a $P < 0.01$. A possible explanation for the fact that the pixel classification outperforms the other methods is the fact that this is the only supervised method (i.e. trained with examples). For a segmentation problem as complicated as the one at hand it is very hard to establish rules which work in all types of situations that can occur in a large set of images.

Choosing a good criterion to measure the performance of vessel segmentation algorithms is not trivial. Whether one method performs better than the other is highly dependent on the application in which the algorithm is to be used. In this study we have used the maximum average accuracy as criterion because our goal was to see which method could most accurately segment the images in the database with respect to the gold

standard. A disadvantage is that in this way the wider vessels have a larger influence on the end result than the smaller vessels. Figure 3 shows the binary result images of each automatic method together with the gold standard and second manual segmentation. The differences between the automated and the manual segmentation are immediately obvious. Many of the smallest vessels of the gold standard and the second manual segmentation are not or just partly visible in the automatic segmentations. In an application where small vessel detection is critical a method with a lower overall accuracy could still be marked the "better" method if it would segment more of the smallest vessels. The optic disk also causes problems for some methods as can be seen in Figure 3. Sometimes this can result in an oversegmentation (eg. Chaudhuri et al. and Martínez-Pérez et al.) or an undersegmentation (eg. Pixel classification) around the boundary of the optic disk. An optic disk segmentation could help solve these problems by excluding the area from the analysis. In the future we hope to be able to assess the performance of the different vessel segmentation methods by evaluating their influence on the performance of a complete retinal disease screening system.

APPENDIX A. WEBSITE

The DRIVE database is freely available on the internet and we invite interested researchers to use this data in their own experiments. Available online are:

1. Both test set and training set.
2. The gold standard as used in this article.
3. Mask images containing the FOV of each image.
4. The results of our pixel classification method.

For more information please refer to the website of the Image Sciences Institute:
`http://www.isi.uu.nl/Research/Databases`

REFERENCES

1. T. Spencer, J. Olson, K. McHardy, P. Sharp, and J. Forrester, "An image-processing strategy for the segmentation and quantification in fluorescein angiograms of the ocular fundus," *Computers and Biomedical Research* **29**, pp. 284–302, 1996.
2. A. Frame, P. Undrill, M. Cree, J. Olson, K. McHardy, P. Sharp, and J. Forrester, "A comparison of computer based classification methods applied to the detection of microaneurysms in ophthalmic fluorescein angiograms," *Computers in Biology and Medicine* **28**, pp. 225–238, 1998.
3. F. Zana and J. Klein, "A multimodal registration algorithm of eye fundus images using vessels detection and Hough transform," *IEEE Transactions on Medical Imaging* **18**(5), pp. 419–428, 1999.
4. K. Goatman, M. Cree, J. Olson, J. Forrester, and P. Sharp, "Automated measurement of microaneurysm turnover," *Investigative Ophthalmology & Visual Science* **44**(12), pp. 5335–5341, 2003.
5. A. Hoover and M. Goldbaum, "Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels," *IEEE Transactions on Medical Imaging* **22**(8), pp. 951–958, 2003.
6. J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Radiology* **174**, pp. 71–74, 2000.
7. M. Heath, K. Bowyer, and D. Kopans, "Current status of the digital database for screening mammography," *Digital Mammography* , pp. 457–460, 1998.
8. A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging* **19**(3), pp. 203–210, 2000.

9. S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of blood vessels in retinal images using two-dimensional matched filters," *IEEE Transactions on Medical Imaging* **8**(3), pp. 263–269, 1989.
10. M. Martínez-Pérez, A. Hughes, A. Stanton, S. Thom, A. Bharath, and K. Parker, "Scale-space analysis for the characterisation of retinal blood vessels," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI'99*, C. Taylor and A. Colchester, eds., pp. 90–97, 1999.
11. F. Zana and J. Klein, "Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation," *IEEE Transactions on Image Processing* **10**(7), pp. 1010–1019, 2001.
12. X. Jiang and D. Mojon, "Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(1), pp. 131–137, 2003.
13. R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley and Sons, New York, 2nd ed., 2001.
14. C. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology* **21**(9), pp. 720–733, 1986.
15. J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics* **33**, pp. 159–174, 1977.