Evaluation of Publicly Available Blood Vessel Segmentation Methods for Retinal Images

Pavel Vostatek[1], Ela Claridge[2], Pauli Fält[3], Markku Hauta-Kasari[3], Hannu Uusitalo[4], and Lasse Lensu[1]

[1]Machine Vision and Pattern Recognition Laboratory
Department of Mathematics and Physics
Lappeenranta University of Technology
PO Box 20, FI-53851 Lappeenranta, Finland
pavel.vostatek@lut.fi
[2]School of Computer Science, University of Birmingham, United Kingdom
[3]School of Computing, University of Eastern Finland
[4]Department of Ophthalmology/SILK, University of Tampere, Finland

Abstract. Retinal blood vessel structure is an important indicator of disorders related to diseases, which has motivated the development of various image segmentation methods for the blood vessels. In this study, two supervised and two unsupervised retinal blood vessel segmentation methods are quantitatively compared by using five publicly available databases with the ground truth for the vessels. The parameters of each method were optimized for each database with the motivation to achieve good segmentation performance for the comparison and study the importance of proper selection of parameter values. The results show that parameter optimization does not significantly improve the segmentation performance of the methods when the original data is used. However, the methods' performance for new data differs significantly. Based on the comparison, Soares method as a supervised approach provided the highest overall accuracy and, thus, the best generalisability. Bankhead and Nguyen methods' performance were close to each other: Bankhead performed better with ARIADB and STARE, whereas Nguyen was better with DRIVE. Sofka method is available only as an executable, and its performance matched the others only with ARIADB.

1 Introduction

Retinal blood vessel structure is an important indicator of disorders such as diabetes, hypertension and cardiovascular disease [7]. To see the changes in the structure, modern retinal imaging modalities enable efficient diagnosis, monitoring and documentation of various conditions. With the current technology, it is possible to produce quantitative information of the signs of eye diseases such as age-related macular degeneration (AMD), diabetic retinopathy (DR) and glaucoma, as well as signs of other abnormalities related to health. When disease screening programs grow wider, the amount of data increases and often

X. Chen, M. K. Garvin, J. Liu, E. Trucco, Y. Xu (Eds.): OMIA 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, Iowa Research Online, pp. 137–144, 2015. Available from: http://ir.uiowa.edu/omia/2015_Proceedings/2015/

makes the manual diagnosis a bottleneck. Consequently, computer aided diagnosis based on automatic and semi-automatic segmentation tools are important in reducing the workload of medical experts.

Automatic and semi-automatic methods enable high-throughput workflows in image-based data handling and permit effective access to blood vessel characteristic information about patient health. An exhaustive review of retinal imaging and its medical implications has been provided by Abràmoff et al. [1]. To enable the characterization of the blood vessels, several approaches have been proposed for segmenting the vessels from retinal images, see [5] and [3] for reviews. A review of general vessel extraction techniques was published by Kirbas et al. in [8]. The vessels can be segmented in both two and three-dimensional medical images using different screening techniques.

In this paper, pixel-wise classification of blood vessels in two-dimensional color fundus images is studied. The paper provides a quantitative comparison of four segmentation methods, two supervised and two unsupervised, with publicly available implementation. Five publicly available retinal image databases with the ground truth for vessels were used for the testing. Parameters of each method were optimized for each database to exploit segmentation potential of the methods and to study the generalisability of the methods.

2 Databases and methods

2.1 Databases provided with blood vessel ground truth

Information of all databases gathered for testing are summarized in Table 1. Number of images, image dimensions, field of view (FOV) angle and diameter, division into classes and percentage of vessel pixels in ground truth are provided.

2.2 Blood vessel segmentation methods and metrics

Method proposed by Soares et al. in [13] (*Soares method*) is a supervised feature-based segmentation algorithm which uses Gabor wavelet filter responses as the classification features. The method parameters are: set of filter sizes (Λ), classifier type (Gaussian mixture model (GMM), K-nearest neighbor (KNN), least mean square error (LMSE)) and number of training samples (n_s). GMM requires a number of Gaussians for modeling vessels and non-vessels (n_{g1}, n_{g2}), and a number of iterations of an expectation maximization (EM) algorithm (n_i). KNN uses the number of nearest neighbors and LMSE is without parameters.

Method proposed by Sofka et al. in [14] (*Sofka method*) is a supervised classification algorithm based on multiscale matched filtering, and confidence and edge measures. Likelihood of a vessel estimation is utilized. The tool is available pretrained as an executable file with no parameters to set.

Method proposed by Bankhead et al. in [2] (*Bankhead method*) is an algorithm based on vessel enhancement using Isotropic undecimated wavelet transform (IUWT)[16] and thresholding based unsupervised segmentation. After the

Table 1: Summary of database information. N_i stands for the number of images and N_{GT} presents the ground truth information: number of experts and percentage of annotated vessel pixels (per expert).

Name, ref.	N_i	FOV [°]	Dimensions FOV ⌀	Subsets	N_{GT}
ARIADB [4]	143	50°	768x576 739 px	AMD (23) Healthy (61) DR (59)	2 (9.6%, 8.5%)
CHASEDB1 [11]	28	30°	1280x960 916 px	Left eye (14) Right eye (14)	2 (10.1%, 9.7%)
DRIVE [15]	40	45°	768x584 649 px	Training (20) Test (20)	2 (12.7%, 12.3%)
HRF [10]	45	60°	3504x2336 3262 px	Healthy (15) DR (15) Glaucoma (15)	1 (9.13%)
STARE [6]	20	35°	605x700 649 px	–	2 (10.3%, 14.8%)

segmentation post-processing is applied to clean the image. Parameters are as follows: set of wavelet levels (Λ) for IUWT, Percentile (p_t) used as the threshold value, sizes of isolated objects and holes (ξ_s, ξ_h) which are removed in post-process.

Method proposed by Nguyen et al. in [9] (*Nguyen method*) is an unsupervised method based on line operators [12]. Vessel contrast is enhanced by filtration of the image by mask of defined size (W) enhancing pixels along lines with different orientations. Multiple filters with varying length of the line ($l_{1..n}$) and linearly combined to produce single image with enhanced vessel contrast. Number of the varying lengths is defined by step ω. Output of the algorithm is a gray-scale map. Thresholding by threshold τ is used to produce binary map.

3 Experiments and results

In order to study the performance of each algorithm, experiments were set up to find the optimal parameters for every tested database (see Section 3.1). Segmentation performance of each algorithm was assessed using the manual segmentation in the databases as a ground truth and accuracy (Acc), sensitivity (Sn) and specificity (Sp) as performance measures [5]. If two sets of manual segmentations were provided in the database, performance of the second observer was assessed and compared to the performance of the automatic methods. Measurement of the performance was always done only on pixels inside the FOV. In the case of databases without a FOV mask (ARIADB, CHASEDB1, STARE), the mask was created.

Table 2: Sampled values of the algorithms' parameters. Superscript over the set of wavelet levels (e.g. $\{1,2\}^{\leq 2}$) stands for subsets of the indicated size (e.g. $\{1,2\}^{\leq 2} = \{\{1\},\{2\},\{1,2\}\}$). ν_i represents such a step value to combine i operator scales.

Soares method		Bankhead method		Nguyen method	
Λ	$\{1, 2..15\}^{\leq 3}$	Λ	$\{1, 2, \cdots, 10\}^{\leq 3}$	W	$\{5, 10, 15, 20, 30, 50, 75\}$
n_s	$\{2 \cdot 10^5\}$	p_t	$\{.08, .1, \cdots, .16\}$	ω	$\{2, 4, 8, \nu_8, \nu_4, \nu_2\}$
n_{g1}	$\{30\}$	ξ_s	$\{0, 50, .., 500\}$	τ	—
n_{g2}	$\{40\}$	ξ_h	$\{0\}$		
n_i	$\{40\}$				

3.1 Experiment setup

Parameters of the algorithms were optimized by exhaustive search through points in a parameter space obtained as a Cartesian product of sets containing sampling values of each parameter. The sets of the parameter values are defined in Table 2. Some remarks about the sampling points follow; in case of Soares method it was preliminarly observed that classification performance was not changing when classifier settings were $\mathrm{n_s} \geq 1.5 \cdot 10^5$, $\mathrm{n_{g1}} \geq 25$, $\mathrm{n_{g2}} \geq 20$, $\mathrm{n_i} \geq 40$ thus these parameters were fixed (values defined in the Table 2). Performance of the Bankhead method was observed to be influenced very little by parameter ξ_h and it was not used in the experiment. In case of Nguyen algorithm the optimal value of τ parameter was optimized using values in the gray-scale filter response and thus was not sampled.

Soares method as a supervised classification needs training examples so subsets of the databases were used as the training data. Number of training images differ for each database; with DRIVE, the dedicated training set was used; with HRF, 15 random images were selected for training; with ARIADB, 30 random images were selected; STARE and CHASEDB1 followed the same procedure – each was randomly divided into two subsets of the same length and each subset was used to train a classifier, during the classification stage each of the classifiers were used to classify images from the training set of the other classifier. The testing procedure was repeated 20 times and reported results are mean values of the classification.

3.2 Results

Results in Tables 3 (Soares method), 4 (Bankhead method) and 5 (Nguyen method) show performace of the methods with different sets of optimized parameters. Comparison of the performance in this section is done using accuracy in order to keep the presentation of the results simple. In Section 3.3 where comparison of the algorithms is provided, sensitivity and specificity of the segmentation are also included.

Table 3: Optimized Soares method performance using $n_{g1} = 30$, $n_{g2} = 40$, $n_s = 2 \cdot 10^5$ and varying Λ sets. Soares settings were as follows: $n_{g1} = 20$, $n_{g2} = 20$, $n_s = 10^6$

Database	Single w. level		Two w. level		Three w. level		Original [13]	
	Λ	Acc	Λ	Acc	Λ	Acc	Λ	Acc
ARIADB	$\{4\}$	0.933	$\{3,4\}$	0.934	$\{2,5,6\}$	0.934	—	—
CHASEDB1	$\{9\}$	0.942	$\{3,9\}$	0.945	$\{3,8,9\}$	0.945	—	—
DRIVE	$\{3\}$	0.945	$\{2,4\}$	0.947	$\{2,3,6\}$	0.947	$\{2,3,4,5\}$	0.947
HRF	$\{10\}$	0.953	$\{5,14\}$	0.956	$\{5,9,10\}$	0.956	—	—
STARE	$\{3\}$	0.944	$\{2,6\}$	0.948	$\{2,3,5\}$	0.948	$\{2,3,4,5\}$	0.949

Table 4: Results of Bankhead method' performance. Parameters providing the best performances on each database are shown.

	Single w. levels				Two w. levels				Original [2]			
Database	Λ	p_t	ξ_h	Acc	Λ	p_t	ξ_h	Acc	Λ	p_t	ξ_h	Acc
ARIADB	{2}	0.12	250	0.937	{2, 3}	0.10	350	0.937	—	—	—	—
CHASEDB1	{3}	0.14	500	0.936	{3, 4}	0.12	500	0.936	—	—	—	—
DRIVE	{2}	0.16	100	0.939	{2, 3}	0.14	250	0.939	{2, 3}	0.15	75	0.937
HRF	{4}	0.08	500	0.953	{3, 4}	0.10	500	0.955	—	—	—	—
STARE	{2}	0.14	200	0.949	{2, 3}	0.12	300	0.949	—	—	—	—

In the case of Soares and Bankhead methods the tables illustrate the performance change when the number of levels change. In case of Nguyen method the results were not influenced by the change of the step parameter assuming that optimal threshold value was used (the case when the threshold value is not optimal is out of scope of this paper) therefore the step value was fixed at $\omega = 2$. Sofka method does not allow to perform the segmentation with various settings therefore the performance results are summarized directly in the comparison section (Section 3.3).

Table 5: Performance of Nguyen algorithm.

	Our results				Original [9]			
Database	W	ω	τ	Acc	W	ω	τ	Acc
ARIADB	15	2	1.20	0.934	—	—	—	—
CHASEDB1	30	2	1.06	0.935	—	—	—	—
DRIVE	15	2	0.70	0.941	15	2	0.53	0.941
HRF	50	2	1.18	0.955	—	—	—	—
STARE	15	2	0.98	0.946	15	2	0.53	0.932

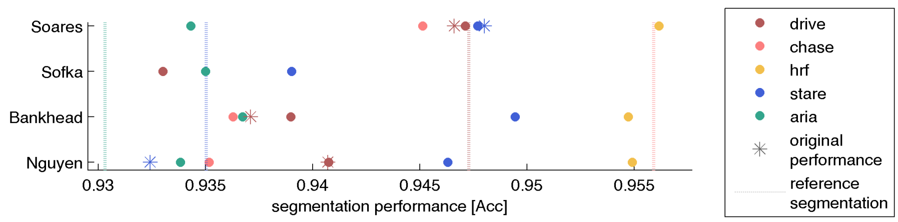

Fig. 1: Comparison of the segmentation performance. The databases are marked with different colors. Solid dots mark the algorithm performance when compared to the ground truth segmentation. Vertical lines mark the performance of second observer when provided. Stars mark the method performance as published in the original papers.

3.3 Comparison of the performance among the algorithms

Table 6 summarizes the optimized performance of each algorithm listed in previous section. Segmentation accuracy, sensitivity and specificity are shown. Of the investigated algorithms the best performance was provided by Soares method providing significantly better performance then other algorithms especially on CHASEDB1 and DRIVE databases. On ARIADB and STARE the differences are not large.

Table 6: Comparison of the optimal results. Appropriate settings for the results are provided in Section 3.2. Results published in original papers are shown in italics under the current results. In order to save space the numbers are expressed as percentages and database names are abbreviated.

	Soares method			Sofka method			Bankhead method			Nguyen method			Reference segm.		
—	Acc	Sn	Sp	Acc	Sn	Sp	Acc	Sn	Sp	Acc	Sn	Sp	Acc	Sn	Sp
AR	93.4	51.3	97.8	93.5	46.9	98.4	93.7	52.5	98.0	93.4	48.9	98.1	93.0	57.6	96.7
CH	94.5	68.8	97.4	92.7	41.3	98.4	93.6	65.1	96.9	93.5	61.3	97.2	95.6	77.0	97.8
DR	94.7	71.4	98.1	93.3	63.0	97.7	93.9	65.4	98.1	94.1	67.8	98.0	94.7	77.6	97.3
	94.7	—	—	—	—	—	*93.7*	*70.2*	*97.2*	*94.1*	—	—	—	—	—
HR	95.6	72.1	98.0	93.9	54.7	97.8	95.5	68.0	98.3	95.5	68.2	98.3		—	—
ST	94.8	70.2	97.7	93.8	53.4	98.4	94.9	63.2	98.6	94.6	66.3	98.0	93.5	89.5	93.8
	94.8	—	—	—	—	—	—	—	—	*93.2*	—	—	—	—	—

Fig. 1 provides an overview of the results and also provides a comparison with the performance of the second observer when provided. With DRIVE, the accuracy of automatic segmentation was similar to the second observer, with CHASEDB1, the second observer was significantly better, and with ARIADB and STARE, the accuracy of second observer was lower than the accuracy of the automatic methods. However, with STARE the reference segmentation yields very high sensitivity while keeping high specificity [13]. Therefore, the performance of the second observer is also better than the automatic methods. Optimization of the results did not provide any improvement in the segmentation performance with the data which were used to develop the methods.

4 Discussion

Five databases used in this study provide range of different image resolutions, images with various pathologies and a broad range of image quality. Comparison to the classification performance of human observers show space to improve segmentation methods mainly on CHASEDB1 and STARE databases where the performance of second observer goes beyond the automatic segmentation.

Most pronounced problems in the resulting segmentations included misclassifications around high contrast segments – around the optic disc, imaging artifacts or pathologies. To some level of extent, supervised classification of Soares eliminates these problems. Sofka method showed good robustness against misclassification in high contrast areas but overall classification performance was poor. All methods also failed often on images with non-even illumination which is expected cause of Soares method failing to perform better on STARE and ARIADB databases while outperforming the unsupervised approaches on DRIVE and CHASE databases.

One more remark is on the parameter optimization process. In order to cover performance in wide range of parameter settings broad span of the searching process was priority and in some cases the performance might be improved by detailed search designed specially for each database. However it is assumed according to observed behavior of the algorithms that no significant performance improvement can be achieved. This is possible line of further development of the experiments.

5 Conclusion

In this study, two supervised (Soares method, Sofka method) and two unsupervised (Bankhead method, Nguyen method) retinal blood vessel segmentation methods were quantitatively compared using five publicly available databases (ARIADB, CHASEDB1, DRIVE, HRF, STARE). The parameters of each method were optimized for each database.

The results show that the parameter optimization does not significantly improve the segmentation performance of the methods when the original data is used. However, the methods' performance for data other than the one or ones used for the development differs significantly. Based on the comparison, Soares method as a supervised algorithm provided the highest overall accuracy and, thus, the best generalisability. Bankhead and Nguyen methods performed comparably: Bankhead method performed better with ARIADB, CHASEDB1 and STARE and Nguyen method better with DRIVE and both methods performed the same with HRF database.

References

1. Abràmoff, M., Garvin, M., Sonka, M.: Retinal imaging and image analysis. IEEE transactions on medical imaging 3, 169–208 (2010)
2. Bankhead, P., Scholfield, C.N., McGeown, J.G., Curtis, T.M.: Fast retinal vessel detection and measurement using wavelets and edge location refinement. PLoS ONE 7(3), e32435 (03 2012), `http://dx.doi.org/10.1371\%2Fjournal.pone.0032435`
3. Bernardes, R., Serranho, P., Lobo, C.: Digital ocular fundus imaging: A review. Ophthalmologica 4, 161–81 (2011)
4. Farnell, D.J., Hatfield, F., Knox, P., Reakes, M., Spencer, S., Parry, D., Harding, S.: Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. Journal of the Franklin institute 345(7), 748–765 (2008)
5. Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A.R., Owen, C.G., Barman, S.A.: Blood vessel segmentation methodologies in retinal images - a survey. Comput. Methods Prog. Biomed. 108(1), 407–433 (Oct 2012), `http://dx.doi.org/10.1016/j.cmpb.2012.03.009`
6. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piece-wise threhsold probing of a matched filter response. IEEE Transactions on Medical Imaging 19(3), 203–210 (March 2000)
7. Kanski, J.J., Bowling, B.: Synopsis of Clinical Ophthalmology. Elsevier Health Sciences, 3rd edn. (2013)
8. Kirbas, C., Quek, F.: A review of vessel extraction techniques and algorithms. ACM Comput. Surv. 36(2), 81–121 (2004)
9. Nguyen, U.T., Bhuiyan, A., Park, L.A., Ramamohanarao, K.: An effective retinal blood vessel segmentation method using multi-scale line detection. Pattern recognition 46(3), 703–715 (2013)
10. Odstrcilik, J., Kolar, R., Budai, A., Hornegger, J., Jan, J., Gazarek, J., Kubena, T., Cernosek, P., Svoboda, O., Angelopoulou, E.: Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database. IET Image Processing 7(4), 373–383 (2013)
11. Owen, C.G., Rudnicka, A.R., Mullen, R., Barman, S.A., Monekosso, D., Whincup, P.H., Ng, J., Paterson, C.: Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program. Investigative ophthalmology & visual science 50(5), 2004–2010 (2009)
12. Ricci, E., Perfetti, R.: Retinal blood vessel segmentation using line operators and support vector classification. Medical Imaging, IEEE Transactions on 26(10), 1357–1365 (2007)
13. Soares, J., Leandro, J., Cesar, R., Jelink, H., Cree, M.: Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. IEEE Transactions on Medical Imaging 25(9), 1214–1222 (September 2006)
14. Sofka, M., Stewart, C.V.: Retinal vessel extraction using multiscale matched filters, confidence and edge measures. IEEE Transactions on Medical Imaging 25(12), 1531–1546 (Dec 2006), `http://www.vision.cs.rpi.edu/vessels/`
15. Staal, J., Abrámoff, M., Niemeijer, M., Viergever, M., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Transactions on Medical Imaging 23(4), 501–509 (April 2004)
16. Starck, J., Fadili, J., Murtagh, F.: The undecimated wavelet decomposition and its reconstruction. IEEE Transactions on Image Processing 16, 297–309 (2007)