A Deep Learning Design for Improving Topology Coherence in Blood Vessel Segmentation

Ricardo J. Araújo[1,2(✉)], Jaime S. Cardoso[1,3], and Hélder P. Oliveira[1,2]

[1] INESC TEC, Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
ricardo.j.araujo@inesctec.pt
[2] Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal
[3] Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

Abstract. The segmentation of blood vessels in medical images has been heavily studied, given its impact in several clinical practices. Deep Learning methods have been applied to supervised segmentation of blood vessels, mainly the retinal ones due to the availability of manual annotations. Despite their success, they typically minimize the Binary Cross Entropy loss, which does not penalize topological mistakes. These errors are relevant in graph-like structures such as blood vessel trees, as a missing segment or an inadequate merging or splitting of branches, may severely change the topology of the network and put at risk the extraction of vessel pathways and their characterization. In this paper, we propose an end-to-end network design comprising a cascade of a typical segmentation network and a Variational Auto-Encoder which, by learning a rich but compact latent space, is able to correct many topological incoherences. Our experiments in three of the most commonly used retinal databases, DRIVE, STARE, and CHASEDB1, show that the proposed model effectively learns representations inducing better segmentations in terms of topology, without hurting the usual pixel-wise metrics. The implementation is available at https://github.com/rjtaraujo/dvae-refiner.

Keywords: Blood vessel segmentation · Deep learning · Topology

1 Introduction

Blood vessel imaging plays a crucial role in several clinical domains, from diagnosis of diseases such as atherosclerosis and aneurysms, to surgery eligibility in organ transplantation, and even surgery planning and guiding. The high volume of data and the large effort required by clinicians to analyse these images led to the need of automatizing the process, such that computer vision methodologies

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11764, pp. 93–101, 2019.
https://doi.org/10.1007/978-3-030-32239-7_11

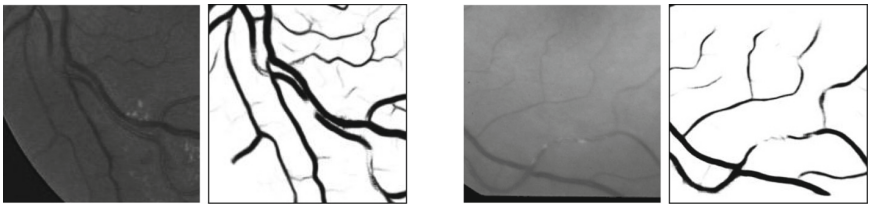

Fig. 1. Example images and corresponding segmentations obtained with the Unet model [8]. Topological errors are common in challenging cases.

started being developed in the end of the past century. Methods relying on strong but intuitive priors quickly advanced the state-of-the-art. With the disposal of labelled databases, essentially containing retinal images and expert delineation of blood vessels, a branch of methodologies using supervision emerged. Lately, with the advent of deep learning, supervised segmentation of blood vessels has reached new performance levels.

Nonetheless, these systems are far from being flawless, as small vessels are still occasionally missed, and the segmented trees often contain topological errors, such as broken vessel segments and sub-segmentation due to central reflex (see Fig. 1). These errors may put at risk applications that require vessel pathways extraction and/or characterization [1], as they may induce relevant differences in the overall blood vessel graph. The state-of-the-art methodologies are prone to commit these topological errors as they rely on minimizing the Binary Cross Entropy (BCE) loss (which only penalizes pixel-wise errors) and mostly use model designs that are not aware of topological incoherences.

Recently, there have been attempts of incorporating topological awareness in deep learning models targeting different applications. A loss encoding hierarchical relations between labels, such as containment and detachment, was designed to improve the multi-class segmentation of histology glands [2]. In [3], a process for consecutive refinement of a segmentation given the grayscale image and previous mask was proposed, guided by the differences between high-level features of the current segmentation and the ground truth. None of these works was applied to vessel segmentation. Bifurcation detection has been addressed in a parallel fashion [4], aiming to enhance the overall segmentation process of vascular networks, and consequently, the overall network topology. Nonetheless, topological errors do not arise in bifurcations only, appearing frequently in the middle of branches due to local loss of signal, as can be seen in Fig. 1.

In this work, we consider adding a refinement step after a typical segmentation network, looking for an end-to-end design that is capable of learning the true vascular topology from noisy and occasionally topological-incoherent observations. To achieve such end, we cascaded a Variational Auto-Encoder (VAE) [5] after a typical segmentation network. We let this VAE take segmentations produced by the latter and reconstruct the ground truth annotation, aiming to learn

a latent space that is capable of avoiding topological incoherences when sufficient evidence is present, leading to topologically coherent segmentations in the end.

2 Methodology

In this section, we present an end-to-end deep neural network design for improving topological consistency in blood vessel segmentation. The methodology comprises a typical segmentation network followed by a refinement model which aims to enforce the learning of meaningful features from corrupted data. We discuss how such design can be used as a strategy to reduce topological mistakes. In what follows, $\mathbf{x}$ and $\mathbf{y}$ denote, respectively, the grayscale input and the ground truth vessel mask, and $\mathbf{y}'$ and $\mathbf{y}''$ represent the outputs of the segmentation network and the refinement model, respectively.

2.1 Auto-Encoding for Learning Local Topology

We can interpret errors committed by the segmentation network as a hidden noise process affecting the true vessel signal $\mathbf{y}$. Thus, we seek an encoding of $\mathbf{y}'$ that does not model this noise, allowing $\mathbf{y}''$ to better depict the topology of $\mathbf{y}$.

Usually, auto-encoding designs encode the entire image $\mathbf{x}$ into a vector $\mathbf{z} \in \mathbb{R}^D$, assuming that complex large scale spatial interactions may be learned. This is not the most adequate option for encoding images where recurring patterns exist, as is the case of blood vessels. In this type of data, for inference purposes, it is better to follow [6], where we have latent variables $\mathbf{z}$ as a 3D tensor (stack of feature maps) instead, for explicitly capturing spatial information.

Let us consider for now the generation process of ground truth vessel masks $\mathbf{y}$. It consists of sampling latent variables from a prior distribution $p_{\theta^*}(\mathbf{z})$ and generating masks according to a conditional distribution $p_{\theta^*}(\mathbf{y}|\mathbf{z})$. We assume these distributions belong to parametric families of distributions $p_\theta(\mathbf{z})$ and $p_\theta(\mathbf{y}|\mathbf{z})$. Given observations $\mathbf{y}$, we want to perform inference in this model, $p_\theta(\mathbf{z}|\mathbf{y}) = \left(p_\theta(\mathbf{y}|\mathbf{z})p_\theta(\mathbf{z})\right)/p_\theta(\mathbf{y})$, to obtain distributions over the latent space explaining the different observations.

The described approach leads to an intractable problem because evaluating the marginal likelihood of the data, $p_\theta(\mathbf{y})$, requires integrating over the entire latent space. This limitation can be circumvented using variational inference by approximating the posterior probability with a family of distributions $q_\lambda(\mathbf{z})$. The optimal parameters λ^* are the ones minimizing the Kullback-Leibler divergence between the two distributions:

$$\begin{aligned} \mathbb{D}_{\mathrm{KL}}(q_\lambda(\mathbf{z})||p_\theta(\mathbf{z}|\mathbf{y})) &= \mathbb{E}_{q_\lambda(\mathbf{z})}\left[\log\left(\frac{q_\lambda(\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{y})}\right)\right] \\ &= \mathbb{E}_{q_\lambda(\mathbf{z})}\left[\log q_\lambda(\mathbf{z}) - \log p_\theta(\mathbf{z},\mathbf{y})\right] + \log p_\theta(\mathbf{y}) \end{aligned} \tag{1}$$

However this optimization problem also requires computing the marginal likelihood, thus being once again intractable. By noting that $\mathbb{D}_{\mathrm{KL}}$ is a non-negative quantity and rearranging (1):

$$\begin{aligned}\log p_\theta(\mathbf{y}) &= \mathbb{D}_{\mathrm{KL}}(q_\lambda(\mathbf{z})||p_\theta(\mathbf{z}|\mathbf{y})) + \mathbb{E}_{q_\lambda(\mathbf{z})}\left[\log p_\theta(\mathbf{z},\mathbf{y}) - \log q_\lambda(\mathbf{z})\right] \\ &\geq \mathbb{E}_{q_\lambda(\mathbf{z})}\left[\log p_\theta(\mathbf{y}|\mathbf{z})\right] - \mathbb{D}_{\mathrm{KL}}\left(q_\lambda(\mathbf{z})||p_\theta(\mathbf{z})\right)\end{aligned} \tag{2}$$

we obtain the Evidence Lower BOund (ELBO), which can equivalently be maximized, allowing us to do approximate posterior inference.

The Variational Auto-Encoder (VAE) [5] conditions the approximate posterior on the data. This distribution, $q_\phi(\mathbf{z}|\mathbf{y})$, and the data likelihood one, $p_\theta(\mathbf{y}|\mathbf{z})$, are both parameterized by neural networks, which are commonly designated as recognition and generative models, respectively. The weights of both networks are jointly learned using the Stochastic Gradient Variational Bayes estimator introduced in the same work. Since parameters ϕ are shared among all observations, this model performs amortized inference.

Until now, we considered the case of auto-encoding the vessel masks $\mathbf{y}$. However, the aim of this work is to use the VAE as a segmentation refiner. Therefore, our recognition model is conditioned by the segmentation output $\mathbf{y}'$, while the generative model produces masks $\mathbf{y}''$ closer to the ground truth $\mathbf{y}$. As we shall discuss next, this formulation is a particular case of a Denoising VAE [7].

2.2 Refinement Model as a Denoising VAE

The Denoising VAE (DVAE) [7] is trained on noisy observations, where the noise is modeled by a distribution conditioned on the data, $p_\gamma(\mathbf{y}'|\mathbf{y})$. In our use case, the outcome of the segmentation network, $\mathbf{y}'$, is interpreted as a corrupted version of the true vessel signal, $\mathbf{y}$. In a DVAE, the recognition model is given by:

$$\tilde{q}_\phi(\mathbf{z}|\mathbf{y}) = \int q_\phi(\mathbf{z}|\mathbf{y}')p_\gamma(\mathbf{y}'|\mathbf{y})d\mathbf{y}' \tag{3}$$

The modified ELBO of the DVAE comes as:

$$\mathbb{E}_{\tilde{q}_\phi(\mathbf{z}|\mathbf{y})}\left[\log\left(\frac{p_\theta(\mathbf{z},\mathbf{y})}{\mathbb{E}_{p_\gamma(\mathbf{y}'|\mathbf{y})}\left[q_\phi(\mathbf{z}|\mathbf{y}')\right]}\right)\right] \tag{4}$$

but a more practical lower bound was proven to be eligible for optimization by the authors [7]:

$$\mathbb{E}_{\tilde{q}_\phi(\mathbf{z}|\mathbf{y})}\left[\log\left(\frac{p_\theta(\mathbf{z},\mathbf{y})}{q_\phi(\mathbf{z}|\mathbf{y}')}\right)\right] \tag{5}$$

which is equivalent to training a regular VAE on corrupted examples. From (5) follows the conclusion that the recognition model in the DVAE is trying to learn meaningful features from the noisy observations, in order to obtain latent representations that allow the generative model to produce a result that is close to the noiseless data. Our proposed refinement model can be seen as a particular case of a DVAE, when the noise model is not known and is encoded in the observations $\mathbf{y}'$. In Fig. 2, we present the proposed design for obtaining segmentation masks that are topologically more coherent.

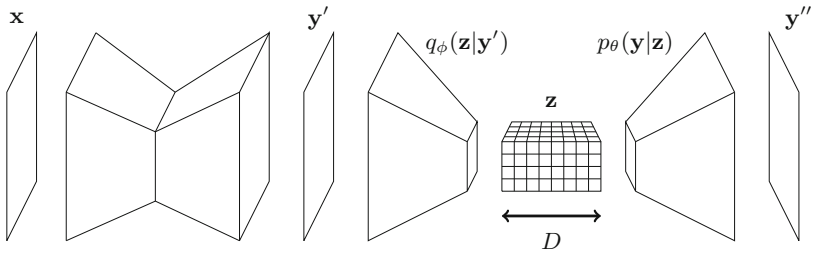

Fig. 2. Design of the proposed model for blood vessel segmentation.

3 Experiments and Discussion

The Unet model [8] is very popular for segmenting biomedical images, given its capability of accounting for both low and high-level features of the images. In this work, the Unet was used as the segmentation network. Our proposed method (*prop*) was compared against two baselines: (i) a single Unet which produces the vessel masks (*unet*), and (ii) a cascade of two Unet models which performs segmentation and refinement tasks (*dunet*). The losses of these models are, respectively, $\mathcal{L}_{prop} = \alpha\mathcal{L}_1(\mathbf{y}', \mathbf{y}) + (1-\alpha)\left(\mathcal{L}_2(\mathbf{y}'', \mathbf{y}) + \mathbb{D}_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{y}')||p_\theta(\mathbf{z}))\right)$, $\mathcal{L}_{unet} = \mathcal{L}_1(\mathbf{y}', \mathbf{y})$ and $\mathcal{L}_{dunet} = \alpha\mathcal{L}_1(\mathbf{y}', \mathbf{y}) + (1-\alpha)\mathcal{L}_2(\mathbf{y}'', \mathbf{y})$, with $p_\theta(\mathbf{z})$ being the standard Gaussian. We tested the impact of using losses $\mathcal{L}_1$ and $\mathcal{L}_2$ other than BCE to train the models: the class-weighted BCE (BCEw), which penalizes more false negatives than false positives (weights of 0.7 and 0.3 were found appropriate for vessel and non-vessel classes); and the focal loss (FL) [9], which is an extension of BCE focusing more on the misclassified examples.

3.1 Datasets

We performed experiments in three of the most used benchmarks for retinal vessel segmentation, DRIVE [10], STARE [11], and CHASEDB1 [12] databases. DRIVE contains some images showing signs of early diabetic retinopathy, half of the images from STARE are pathological, and CHASEDB1 comprises data where central vessel reflex is abundant. Only DRIVE splits by default the images into train and test sets, each with 20 images. In this work, we set apart the last 10 images of STARE and CHASEDB1 for testing purposes, and used the remaining for training the models.

3.2 Metrics

To compare the performance of the models, we considered usual pixel-wise metrics such as: the area under the ROC curve (AUC), sensitivity, and specificity. To evaluate the topological coherence of the masks, we followed a similar approach to [13]. A connected path is randomly chosen from the ground truth and the equivalent path in the binarized prediction mask is analysed. The prediction

is classified as infeasible if such path does not exist. Otherwise, it is wrong or correct whether its length differs by more than 10%, or not, respectively. We sampled 1000 paths per test image.

3.3 Implementation Details

The train data of DRIVE was randomly split into 15 training and 5 validation images, in order to tune the models. When these included a refinement step, α was set to 1 and decreased 5e-3 each epoch until 0.3, and $\mathcal{L}_1 = \mathcal{L}_2$. For stability purposes, when using the focal loss, we set $\mathcal{L}_1 = \text{BCE}$ and $\mathcal{L}_2 = \text{FL}$. The training procedure lasted for 150 epochs where, in each, 300 batches of 16 patches of size 64×64 were used. Patches were taken from the green channel of images and augmented via random transformations including horizontal and vertical flips, rotations in the range $\left[-\frac{\pi}{2}, +\frac{\pi}{2}\right]$, and addition of an intensity bias.

The original Unet model comprises 4 condensing and expanding levels, however we concluded that 2 were ideal in this scenario, when considering the AUC metric. Afterwards, we tuned the refiners in the Double Unet and proposed models, considering the number of correct paths. Both pipelines ended having around 4M parameters. The best performing Double Unet model was a cascade of two Unets as described above. Our proposed recognition model was constituted by 4 convolutional layers (3×3 kernels and padding of 1) producing, respectively, 64, 64, 256, and 256 feature maps. Each of the first 3 is followed by a max pooling layer (kernel size of 2). Then, convolutional layers (1×1 kernels, no padding) learning D feature maps, parameterize the diagonal Gaussian over the latent space. D was tuned to 100. Regarding the generative model, it includes 3 transposed convolutional layers (4×4 kernels, padding and stride of 2) producing, respectively, 256, 256, and 64 feature maps, followed by 2 convolutional layers (3×3 kernels and padding of 1), where the first learns 64 kernels and the last outputs the parameters of a Bernoulli distribution. ReLUs were used in the intermediate layers of the proposed VAE, and a Sigmoid activation function in the last one. The modulating constant of $\mathbb{D}_{\text{KL}}$ was tuned to 1e-3.

Having tuned the structure and hyperparameters of the models, they were trained as before, but this time using all the train data. Note that we perform patch-based training, but the design of the models allows single-pass prediction.

3.4 Results and Discussion

The average performance of the models on 5 different runs is shown in Table 1. The focal loss slightly increased the AUC of the models, meaning that they became better at separating both classes. However, that did not necessarily translate into better topological masks in the end. This is not surprising, as giving more focus to hard cases does not guarantee we are giving more weight to the pixels that generate topological mistakes. Instead, using BCEw, thus giving more weight to the vessel class, allowed to improve the sensitivity and the topology, as was expected. Proceeding to model design comparison, the proposed method was able to significantly decrease the number of infeasible paths, essentially

Table 1. Performance of the models, in percentage, averaged over 5 runs. *AUC*, *sen*, *spe*, *inf*, and *cor* stand for, respectively, area under the roc curve, sensitivity, specificity, infeasible, and correct paths. Note that lower *inf* values are better.

		BCE			BCEw			FL		
		unet	dunet	prop	unet	dunet	prop	unet	dunet	prop
DRIVE	*AUC*	97.7	97.8	97.8	97.9	97.9	97.9	**98.0**	**98.0**	97.9
	sen	79.2	79.6	85.1	87.4	87.8	89.7	78.4	79.0	82.3
	spe	98.1	98.0	96.7	96.2	96.1	95.3	98.1	98.1	97.4
	inf	47.0	45.0	34.1	34.8	31.8	**29.1**	48.6	44.8	40.4
	cor	45.5	47.3	56.7	56.7	59.2	**61.2**	43.8	48.3	51.4
STARE	*AUC*	98.0	98.2	98.3	98.1	98.4	98.6	98.7	**98.8**	**98.8**
	sen	80.5	82.7	87.3	87.7	89.1	90.1	81.1	82.7	85.2
	spe	98.5	98.4	97.3	97.3	97.2	96.8	98.5	98.4	97.9
	inf	53.4	43.2	27.9	38.9	29.2	**23.1**	49.7	38.4	34.9
	cor	40.8	51.8	61.9	55.3	64.3	**69.2**	43.6	54.4	58.1
CHASE	*AUC*	97.6	97.7	97.9	97.8	98.0	98.0	97.9	**98.2**	**98.2**
	sen	80.7	80.5	82.8	87.8	88.4	89.8	80.6	80.9	84.2
	spe	97.6	97.6	97.4	95.9	95.9	95.6	97.5	97.7	97.2
	inf	74.9	74.0	64.7	60.5	54.6	**48.0**	73.7	71.4	62.2
	cor	20.9	22.8	29.8	32.4	38.1	**45.6**	21.6	24.5	31.8

due to finding the correct topology, as demonstrated by the increase of correct paths. This was achieved without hurting pixel-wise metrics, as may be seen by analysing the AUC. In fact, this metric was even improved in some cases. Note that there is a compromise between the sensitivity and specificity of the

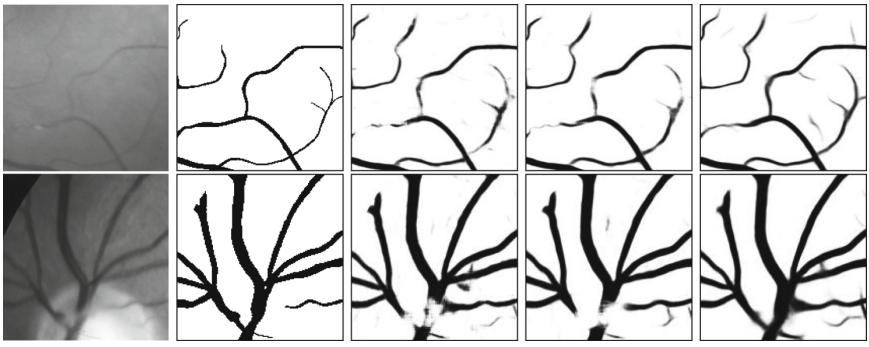

Fig. 3. Example masks obtained by using the BCEw loss. From left to right: original images, ground truth, and predictions from Unet, Double Unet, and proposed method.

models, such that using them for direct comparison of models is often not trivial. By comparing against the Double Unet results, which is also a model with more capacity, we conclude that our proposed design effectively learned better features for ensuring topological coherence. Figure 3 shows some visual results of the three models trained with BCEw.

4 Conclusion

We proposed a design where a VAE is cascaded after a segmentation network, with the purpose of improving the topological coherence of the predicted blood vessel masks. The experiments showed that our methodology achieves that objective by predicting more correct paths and less infeasible paths, without negatively affecting pixel-wise metrics. The results of comparing the proposed method with a cascade of two Unet models sustain that the improvement comes from the model design and not from the increased complexity of the pipeline. Future works will include investigating a differentiable loss that is aware of the topology. Besides, we believe the design here proposed may be useful for semi-supervised segmentation of blood vessels.

Acknowledgements. This work was financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within PhD grant number SFRH/BD/126224/2016 and within project UID/EEA/50014/2019.

References

1. Zhao, Y., et al.: Retinal artery and vein classification via dominant sets clustering-based vascular topology estimation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 56–64. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_7
2. BenTaieb, A., Hamarneh, G.: Topology aware fully convolutional networks for histology gland segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 460–468. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_53
3. Mosinska, A., Marquez-Neila, P., Koziński, M., Fua, P.: Beyond the pixel-wise loss for topology-aware delineation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3136–3145. IEEE, Salt Lake City (2018)
4. Uslu, F., Bharath, A.A.: A multi-task network to detect junctions in retinal vasculature. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 92–100. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_11
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Proceedings of International Conference on Learning Representations 2014, Banff (2014)
6. Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improved variational inference with inverse autoregressive flow. In: Advances in Neural Information Processing Systems, vol. 29, pp. 4743–4751. Curran Associates Inc., Barcelona (2016)

7. Im, D.I.J., Ahn, S., Memisevic, R., Bengio, Y.: Denoising criterion for variational auto-encoding framework. In: Thirty-First AAAI Conference on Artificial Intelligence. AAAI Press, San Francisco (2017)
8. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
9. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988. IEEE, Venice (2017)
10. Staal, J., Abrámoff, M.D., Niemeijer, M., Viergever, M.A., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)
11. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. Imaging **19**(3), 203–210 (2000)
12. Owen, C.G., et al.: Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program. Investig. Ophthalmol. Vis. Sci. **50**(5), 2004–2010 (2009)
13. Wegner, J.D., Montoya-Zegarra, J.A., Schindler, K.: A higher-order CRF model for road network extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1698–1705. IEEE, Portland (2013)