

University of Dundee

DOCTOR OF PHILOSOPHY

**Leveraging Modelling and Machine Learning for the Analysis of Curvilinear Structures
in Medical Images**

Annunziata, Roberto

Award date:
2016

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Leveraging Modelling and Machine Learning for the Analysis of Curvilinear Structures in Medical Images



Roberto Annunziata

*University of Dundee,
United Kingdom*

This dissertation is submitted in partial fulfilment for the degree of
Doctor of Philosophy

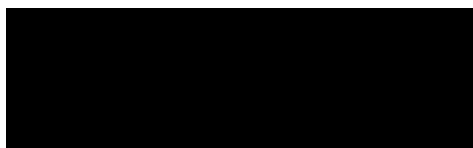
*School of Science and
Engineering (Computing)*

August 2016

Declaration of Authorship

Candidate's declaration

I, Roberto Annunziata, hereby declare that I am the author of this thesis; that I have consulted all references cited herein; that the work of which this thesis is a record has been done by me, and that it has not been previously accepted for a higher degree.



Signed

18/8/2016

Date

Supervisor's declaration

I, Emanuele Trucco, hereby declare that I am the supervisor of the candidate, and that the conditions of the relevant Ordinance and Regulations have been fulfilled.

Signed

18/8/2016

Date

To my mother

*the person who made me realize
the importance of being willing to learn*

Acknowledgements

I am grateful to my supervisor, Prof. Emanuele Trucco, for his continued support, guidance and encouragement during these wonderful 3 years. I appreciated his constructive criticism to my ideas and the way I approached scientific problems, which helped me strengthen both my technical and presentation skills. In particular, I acknowledge him for acting as the picky reviewer (more on my worst-case scenario below), and for setting the bar always high.

I would like to thank the clinical collaborators for my project, Prof. Dr Pedram Hamrah (Harvard Medical School/Tufts Medical Center, USA) and Dr Ahmad Kheirkhah (Harvard Medical School, USA), for providing a large quantity of image data, annotations, suggestions and for hosting me there in Boston for an unforgettable 2-month placement.

Thanks to the CVIP and VAMPIRE groups (University of Dundee, UK), for all the fruitful discussions and for innumerable conversations during lunch, which let me learn a lot more about other amazing cultures.

I want to express my gratitude to Prof. Stephen J. McKenna (CVIP), always available to provide constructive feedback, and to Dr Jianguo Zhang (CVIP) for having successfully acted as the worst-case scenario in terms of potential reviewers.

This research was financially made possible by the European Union Marie-Sklodowska Curie Initial Training Network (ITN) "REtinal VAscular Modelling, Measurement And Diagnosis" (REVAMMAD), project number 316990. I wish to thank Prof. Andrew Hunter (University of Lincoln, UK) for having led the whole project with passion and enthusiasm, and for valuable advice.

Finally, a special thank goes to my mother and my girlfriend for their continued moral support, and for letting me focus on this research.

List of Publications and Achievements

1 List of publications

1.1 Journals

- **R. Annunziata** and E. Trucco, "Accelerating Convolutional Sparse Coding for Curvilinear Structures Segmentation by Refining SCIRD-TS Filter Banks", *IEEE Transactions on Medical Imaging* (in press). [e-version].
- **R. Annunziata**, A. Kheirkhah, S. Aggarwal, P. Hamrah and E. Trucco, "A Fully Automated Tortuosity Quantification System with Application to Corneal Nerve Fibres in Confocal Microscopy Images", *Medical Image Analysis*, Vol. 32, pp 216-232, April 2016. [e-version]
- **R. Annunziata**, A. Kheirkhah, S. Aggarwal, B. M. Cavalcanti, P. Hamrah and E. Trucco, "Two-Dimensional Plane for Multi-Scale Quantification of Corneal Subbasal Nerve Tortuosity", *Investigative Ophthalmology & Visual Science*, Vol. 57, pp 1132-1139, March 2016. [e-version].
- **R. Annunziata**, A. Garzelli, L. Ballerini, A. Mecocci and E. Trucco, "Leveraging Multiscale Hessian-based Enhancement with a Novel Exudate Inpainting Technique for Retinal Vessel Segmentation", *IEEE Journal of Biomedical and Health Informatics* (in press). [e-version].

1.2 Refereed conference papers

- **R. Annunziata**, A. Kheirkhah, P. Hamrah and E. Trucco, "Scale and Curvature Invariant Ridge Detector for Tortuous and Fragmented Structures", *Proc. of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Part III, pp 588-595, Munich, DE, October 5-9, 2015. [Acceptance rate: 32.5%] [e-version].
- **R. Annunziata**, A. Kheirkhah, P. Hamrah and E. Trucco, "Boosting Hand-Crafted Features for Curvilinear Structure Segmentation by Learning Context Filters", *Proc. of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Part III, pp 596-603, Munich, DE, October 5-9, 2015. [Acceptance rate: 32.5%] [e-version].

- **R. Annunziata**, A. Kheirkhah, S. Aggarwal, B. M. Cavalcanti, P. Hamrah and E. Trucco, "Tortuosity Classification of Corneal Nerves Images Using a Multiple-Scale-Multiple-Window Approach",
Proc. of the Ophthalmic Medical Image Analysis First International Workshop, OMIA 2014, – MICCAI 2014, p.113-120, Boston, USA, September 14, 2014. [e-version].
- **R. Annunziata**, A. Kheirkhah, P. Hamrah and E. Trucco, "Combining Efficient Hand-Crafted Features with Learned Filters for Fast and Accurate Corneal Nerve Fibre Centreline Detection",
Proc. of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society – EMBC 2015, pp 5655 - 5658, Milan, IT, August 25-29, 2015. [Oral Presentation] [e-version].
- A. Lisowska, **R. Annunziata**, E. Trucco, D. Karl, G.K. Loh, "An Experimental Assessment of Five Indices of Retinal Vessel Tortuosity with the RET-TORT Public Dataset",
Proc. of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society – EMBC 2014, pp 5414 - 5417, Chicago, USA, August 26-30, 2014. [e-version].
- **R. Annunziata**, A. Kheirkhah, P. Hamrah and E. Trucco, "Towards Robust and Efficient Automated Curvilinear Structure Detection in Medical Images",
PhD symposium 2015, School of Computing, University of Dundee, Dundee, UK. [Best Paper Award]

1.3 Unrefereed workshop papers

- S. Manivannan, H. Shen, W. Li, **R. Annunziata**, H. Hamad, R. Wang and J. Zhang, "Brain Tumor Region Segmentation using Local Co-occurrence Features and Conditional Random Fields",
Proc. of Digital Pathology Classification and Segmentation Challenge, MICCAI, 2014, Boston, USA.
- E. Trucco, **R. Annunziata**, L. Ballerini, E. Pellegrini, S. McKenna, T. MacGillivray, T. Pearson, D. Relan, G. Robertson, A. Doney, B. Dhillon and J. Cameron, "VAMPIRE: Vessel Assessment and Measurement Platform for Images of the RETina",
SICSA Dundee Medical Image Analysis Workshop, 2015, Dundee, UK.

- **R. Annunziata**, A. Kheirkhah, S. Aggarwal, B.M. Cavalcanti, P. Hamrah and E. Trucco, "A Supervised Approach for Classifying Corneal Nerves Images in 4 Tortuosity Levels",
6th SINAPSE Annual Scientific Meeting and SPIRIT Workshop, 2014, Edinburgh, UK.
- T. MacGillivray, G. Robertson, D. Relan, E. Pellegrini, K. Zutis, **R. Annunziata**, L. Ballerini, J. Cameron, B. Dhillon, E. Trucco, C. Lupascu, D. Tegolo, A. Giachetti, A. Doney and P. Wilson, "VAMPIRE: Vasculature Assessment and Measurement Platform for Images of the RETina",
6th SINAPSE Annual Scientific Meeting and SPIRIT Workshop, 2014, Edinburgh, UK.

2 Achievements

- ***MICCAI Travel Award 2015.***

The MICCAI Student Travel Awards scheme rewards the best (i.e. highest scoring) first author students and subsidises their attendance to present their work at the annual MICCAI conference.

- ***Best Paper Award for Outstanding Doctoral Student***

PhD Symposium 2015, School of Computing, University of Dundee, Dundee, UK.

- ***Winner of the REVAMMAD Machine Learning Challenge: “Classification of image patches into lesion/non-lesion”***

Contest held in conjunction with a training event within the REVAMMAD project (EU ITN in the 7th framework of the Marie Skłodowska-Curie Actions), Hersonissos, Crete, EL, June 2014.

- ***Member of the runner-up team for the international “Brain Tumor Digital Pathology Segmentation Challenge”***

MICCAI 2014, Boston, USA.

Training Activities and Secondments

1 Training

1.1 Medical Imaging Summer School 2014 (MISS2014), Favignana, IT

This summer school was endorsed by the MICCAI society and was organised by Prof. R. Cipolla (Univ. of Cambridge, UK), Prof. J. Schnabel (Univ. of Oxford), Prof. G. M. Farinella and F. Stanco (Univ. of Catania, IT).

Poster Presentation: "A Novel Supervised Method for Corneal Fibres Tortuosity Classification Using a Multi-Scale-Multi-Window Approach"

1.2 REVAMMAD training events

The REVAMMAD project was part of European Union's 7th Framework (FP7) Marie Curie Initial Training Network programme; as such, it included a number of training events and related activities:

- **Padova, IT (summer 2015):**
 - Topics: advanced modelling of the eye vascular system, statistics, optics, EU funding.
 - Presentation: current progress.
- **Berlin, DE (winter 2014):**
 - Mid-term review meeting;
 - Topics: research skills, clinical factors for screening, eye diseases, writing and presenting research results, biological experiments.
 - Presentation: current progress delivered to EU representatives.
- **Crete, EL (summer 2014):**
 - Topics: machine learning, advanced vascular image processing, micro vascular modelling, lesion detection, active contours, ESR symposium, media training.
 - Presentations: (1) current progress, (2) vascular adaptive responses: tortuosity.
- **Lincoln, UK (winter 2013):**
 - Topics: clinical background on the eye, vascular modelling, image processing, research skills, brainstorming.

- Presentation: Background and current progress.

2 Secondments

As REVAMMAD early stage researcher I had the possibility to spend up to 7 months at partner institutions and companies and I took full advantage of this chance. Specifically, I spent a secondment period in the following groups:

- Prof. Dr P. Hamrah, Dr. A. Kheirkhah and their research group at Harvard Medical School (MEEI) and Tufts Medical Center, Boston, USA;
- Prof. Dr A. Pries, Dr B. Reglin and their research group at Charité - Universitätsmedizin Berlin, DE.
- Dr J. van Hemert, Dr A. Fleming and their research group at OPTOS plc (NIKON), Edinburgh, UK.

Abstract

Recent clinical research has highlighted important links between a number of diseases and the tortuosity of curvilinear anatomical structures such as corneal nerve fibres, suggesting that tortuosity changes might detect early stages of specific conditions. Currently, clinical studies are mainly based on subjective, visual assessment, with limited repeatability and inter-observer agreement.

In this thesis I have endeavoured to address these problems by proposing a fully automated framework for image-level tortuosity estimation, consisting of a hybrid segmentation method and a versatile tortuosity estimation algorithm. The former combines an *appearance* model, based on a *Scale and Curvature Invariant Ridge Detector* (SCIRD), with a *context* model, including multi-range learned context filters. The latter is based on a novel tortuosity estimation paradigm in which discriminative, *multi-scale* features can be automatically learned for specific anatomical objects and diseases.

I have validated each module of the system separately and then assessed their impact on the tortuosity estimation performance (target application). The segmentation module has been tested on 5 challenging data sets, including corneal nerve fibres (not public, provided by our clinical collaborators at MEEI, Harvard Medical School, USA), neurites (2 benchmark data sets) and retinal blood vessels (2 benchmark data sets). The tortuosity estimation module has been validated on a data set including 140 corneal nerve images, the largest ever used for this task, to my best knowledge.

Experimental results show that (1) the segmentation module outperforms state-of-the-art hand-crafted and hybrid approaches; (2) the tortuosity estimation module performs better than state-of-the-art and widely used tortuosity indices; (3) the whole system matches and sometimes even exceeds tortuosity estimation performance of experienced observers when compared against each other, a level of performance that will allow us to deploy the system on much larger data sets, with the aim of discovering new links between tortuosity and specific diseases in an objective and repeatable fashion.

Contents

Contents	xii
List of Figures	xvi
List of Tables	xxi
1 Introduction	1
1.1 Background and motivation: curvilinear structure analysis in medical images	1
1.2 Contributions	4
1.2.1 Novel curvilinear structure ridge detectors	4
1.2.2 Learning single- and multi-range context filters	5
1.2.3 Accelerating convolutional sparse coding for filter learning	5
1.2.4 Multi-scale analysis of tortuosity measures	5
1.2.5 A machine learning approach to tortuosity definition	6
1.3 Thesis organisation	6
2 Related Work	9
2.1 Introduction	9
2.2 Curvilinear structure modelling and segmentation	11
2.2.1 Hand-crafted filters	12
2.2.2 Fully learned architectures	12
2.2.3 Hybrid methods	13
2.3 Tortuosity quantification	14
2.3.1 Corneal nerve fibres tortuosity assessment	17
2.4 Discussion	17
2.4.1 Limitations: Modelling and Segmentation	18
2.4.2 Limitations: Tortuosity estimation	19
2.5 Conclusions	20

3	Materials and Experimental Protocol	22
3.1	Introduction	22
3.2	Data sets	22
3.2.1	IVCM100	22
3.2.2	IVCM140	24
3.2.3	BF2D	24
3.2.4	VC6	25
3.2.5	DRIVE	26
3.2.6	STARE	26
3.3	Performance evaluation protocol	29
3.3.1	Segmentation	29
3.3.2	Tortuosity estimation	30
3.4	Conclusions	31
4	Curvilinear Structure Modelling and Segmentation	33
4.1	Introduction	33
4.2	SCIRD: a Scale and Curvature-Invariant Ridge Detector	34
4.2.1	Curvilinear structure model	34
4.2.2	SCIRD	36
4.2.3	Supervised SCIRD	40
4.2.4	Experiments and results	42
4.3	Learning context filters	45
4.3.1	Unsupervised filter learning	47
4.3.2	Context filters	49
4.3.3	Description vector and supervised classification	50
4.3.4	Experiments and results	52
4.4	Learning <i>multi-range</i> context filters	55
4.4.1	<i>Multi-range</i> context filters	57
4.4.2	Experiments and results	57
4.5	Conclusions	60
5	Tortuosity Estimation	63
5.1	Introduction	63
5.2	A machine learning approach to tortuosity estimation	63
5.2.1	Multi-scale tortuosity representation of a single curvilinear structure	65
5.2.2	Image-level tortuosity features for image classification	68
5.2.3	Feature selection and tortuosity prediction	69

5.2.4	Experiments and results	70
5.3	Tortuosity plane and confidence of the estimated tortuosity level	77
5.4	Conclusions	80
6	Improving Curvilinear Structure Modelling and Segmentation	83
6.1	Introduction	83
6.2	Improving SCIRD	84
6.2.1	SCIRD for thin structures (SCIRD-TS)	84
6.2.2	Experiments and results	88
6.3	Improving unsupervised filter learning	89
6.3.1	Optimal warm-start strategy	93
6.3.2	Refining the prototype filters by CSC	95
6.3.3	Impact of the warm-start strategy on CSC optimisation	95
6.3.4	Experiments and results	97
6.4	Conclusions	104
7	Conclusions, Discussion and Future Work	109
7.1	Introduction	109
7.2	Summary of the thesis	109
7.2.1	Curvilinear structure modelling and segmentation	111
7.2.2	Tortuosity estimation	112
7.2.3	Improving curvilinear structure modelling and segmentation	112
7.3	Contributions	113
7.4	Limitations of the proposed tortuosity estimation system and future work	114
7.4.1	Segmentation module	114
7.4.2	Tortuosity estimation module	115
7.4.3	Interpretation of tortuosity estimates	115
7.4.4	Experiments	116
	Bibliography	117
	Appendix A Further Work Carried Out During the Project	128
A.1	Investigating the biological factors generating tortuosity	128
A.1.1	Related work	128
A.1.2	Materials	129
A.1.3	Methods	131
A.1.4	Experiments and results	132

A.1.5 Conclusions 133

List of Figures

1.1	Examples of curvilinear structures considered. (a) Neurites; (b) retinal blood vessels; and (c) corneal nerve fibres.	2
1.2	Thesis roadmap: chapters and their relation.	8
2.1	Factors making curvilinear structure segmentation challenging. (a) a corneal nerve fibre with low signal-to-noise ratio; (b) a fragmented neurite; (c) retinal blood vessels surrounded by exudates (i.e., confounding non-target structures); (d) corneal nerve image with non-uniform illumination; and (e) corneal nerve fibres with complex configurations (e.g., tortuosity, bifurcations and parallel fibres).	11
2.2	(a) Key quantities for the DM; (b) vessel partitioning used by the TD algorithm.	15
3.1	Four examples of corneal nerve fibre images captured through <i>in vivo</i> confocal microscopy. A,B,C,D: original images with increasing tortuosity level (A = 1, D = 4). E, F, G, H: manually traced fibre centrelines for image A, B, C, D, respectively.	23
3.2	Four examples of corneal nerve fibre images from subjects with Herpes Simplex Virus keratitis.	24
3.3	Original images (first row) and ground truth segmentation (second row) from the BF2D data set.	25
3.4	Original images (first row) and ground truth segmentation (second row) from the VC6 data set.	26
3.5	Original images (first row) and ground truth segmentation (second row) from the DRIVE data set.	27
3.6	Original images (first row) and ground truth segmentation (second row) from the STARE data set.	28

4.1	The need for a curved-support model. (Left) 3-D profile of the tortuous corneal nerve fibre highlighted with the red rectangle in the original image (right).	34
4.2	<i>Appearance</i> model used to capture specific characteristics of tortuous and fragmented structures. (a) An example of our curved-support shape model; (b) the insert (below) visualises the direction along which we measure the second (directional) derivative used to compute the <i>tubularity probe filter</i> of the shape model (above), in the window selected; (c) Derived convolutional filter, estimating local tubularity efficiently.	36
4.3	The effect of geometric parameters on the shape of SCIRD filters.	38
4.4	(a) A subset of SCIRD filters used in our experiments. Notice, the model includes straight support too; (b) original images (top), supervised SCIRD <i>tubularity</i> maps (bottom) from IVCN (left), BF2D (centre), VC6 (right) testing sets.	41
4.5	Qualitative comparison: tubularity estimation results on IVCN. SCIRD (our approach) shows better connectivity and higher signal-to-noise ratio than others.	42
4.6	Precision-recall curves for SCIRD and baselines on IVCN, BF2D and VC6 datasets. Curves are obtained applying different thresholds on the <i>tubularity</i> maps.	44
4.7	Visual examples motivating the need for including <i>context</i> information in the segmentation pipeline. (a - blue circles) short, well contrasted, but isolated segments, which could be wrongly segmented as corneal nerve fibres (they are dendritic cells, instead); and (b - red circles) short and poorly contrasted segments branching from corneal nerve fibres, which should be segmented as corneal nerve fibres.	46
4.8	Difference in learning appearance and context filters. (Left) Appearance filters are learned using original image patches and applied on the same layer of the HCFs, thus leading to potential redundancy in the feature set. (Right) Context filters are learned using HCF maps and applied after the HCFs, thus eliminating redundancy.	48
4.9	Original images (top) and tubularity maps (bottom) obtained with our approach on IVCN (left), BF2D (centre), VC6 (right) datasets.	51
4.10	Precision-recall curves for pixel-level classification. Shaded color bands represent 1 standard deviation of the results from individual runs.	53

4.11	Block diagram of the proposed unsupervised <i>multi-range</i> context filters learning. Notice, we keep the patch size constant over the levels to capture larger and larger spatial context (light orange, orange and red indicate short-, medium- and long-range context, respectively).	56
4.12	Precision-recall curves for the corneal nerve fibre centreline detection task (pixel-level classification). Each curve is obtained by averaging the results of 5 random cross-validation trials in which the whole data set, including 140 images, is randomly split in 2 equal partitions.	59
4.13	Examples from segmentation experiments on IVCMI40. First row: original images. From second to fourth row: probability maps obtained by combining OOF with learned <i>appearance</i> filters [111], combining OOF with <i>single-range</i> context filters and combining SCIRD with <i>multi-range</i> learned context filters (proposed approach). Last row: ground truth.	62
5.1	Example of multiple spatial scales observed in a corneal nerve fibre (a, yellow arrows). Multi-scale decomposition (c,d) of a simple model of corneal nerve fibre obtained as sum of two sine waves with different frequencies (b).	64
5.2	Block diagram of our tortuosity estimation framework: wrapper-based feature selection followed by multinomial logistic ordinal regression. First, for every corneal nerve fibre segment detected in an image, robust spline fitting is applied for fast and accurate curvature estimation at each level of the fibre's scale-space representation. Mean curvature $k_{mean}(t)$, twistedness $d_{ip}(t)$, and maximum curvature $k_M(t)$ are computed for each scale $t \in \{1, \dots, t_M\}$ and the related feature vector v_f is built. Second, weighted (fibre length is used as weight) averaging across all the fibres creates the pool of image-level features v_{img} . Third, a wrapper-based FS technique is employed for identifying the most discriminative combination of features and their scale. Finally, a multinomial logistic ordinal regressor is used to assign image-level tortuosity.	66
5.3	Accurate multi-scale curvature estimation via cubic spline fitting. Stem plots illustrate the estimated local curvature for the automatically segmented corneal nerve fibre shown in the top image. Left, right: estimated curvature at scale 2 and 5, respectively. The green arrows indicate starting point ($l = 0$) and direction ($l > 0$).	67

5.4	Multi-scale analysis of a corneal nerve fibre. From left to right: original fibre including turns at all frequencies (spatial scale 1), and its smoothed versions at spatial scales from 2 to 6 to take into account turns at intermediate and high frequencies.	69
5.5	Tortuosity estimated for 4 corneal nerve images in IVCMI40 by the proposed method. Images are projected onto the tortuosity plane as points (markers indicate the tortuosity level assigned by the expert observers) whose coordinates are the estimated mean curvature at scale 2 (high frequency turns) and scale 5 (low frequency turns). The level of confidence for each tortuosity estimate is encoded with colour (red and blue mean high and low confidence, respectively). Intuitively, the confidence is related to the distance from the decision boundaries (indicated in black) separating each tortuosity region.	78
5.6	All images in IVCMI40 are projected onto the tortuosity plane after a leave-one-out cross-validation on unseen images. Markers indicate the majority voting ground truth tortuosity level. The estimated tortuosity level corresponds to the region within which an IVCMI image is mapped. During each cross-validation the best decision boundaries (shown in black) are computed based on the training set.	80
6.1	First row: ideal thin structure (1 pixel wide); second row, from left to right: SCIRD filter, SCIRD response and its cross-sectional profile along the blue line; third row, from left to right: SCIRD-TS filter, SCIRD-TS response and its cross-sectional profile along the blue line. Notice that while the SCIRD response is approximately 0 on the thin structure (i.e. SCIRD does not detect it), the SCIRD-TS one is maximum, hence leading to a correct detection. . .	86
6.2	Detecting thin vessels. Left: original image patch (green channel) showing thin retinal blood vessels around the fovea; middle: enhancement using SCIRD; right: enhancement using the proposed SCIRD-TS. The thin vessels not enhanced by SCIRD are correctly enhanced by SCIRD-TS.	87
6.3	Performance evaluation in terms of precision-recall curves (pixel-level segmentation) for several HCFs on BF2D, DRIVE, STARE and VC6 data sets.	90
6.4	A filter bank learned using convolutional sparse coding with random initialisation. The DRIVE data set was used for this experiment.	92

6.5	Block diagram of the proposed method. Notice that all the curvilinear structures, represented in the space \mathcal{S} as blue and red dots, are copied in the other spaces as well (to show their original position), and they are shown in light blue and light red, respectively.	93
6.6	Experiments: reconstruction error and time to convergence. Performance evaluation in terms of total reconstruction error for CSC with random, DCT and SCIRD-TS initialisation. Each row shows the influence of the dictionary size on the total reconstruction error, for each data set. Optimisations were stopped at convergence.	105
6.7	Visualisation of a CSC-refined SCIRD-TS filter bank. SCIRD-TS filter banks obtained after the fast warm-start strategy (first column), refinement by CSC (second column) and difference (third column) for DRIVE, BF2D, VC6 and STARE (refer to Table 6.1 for time to convergence). Some of the original filters are unchanged, while most of the others are only modified in length or width.	106
6.8	Experiments: segmentation. Performance evaluation in terms of precision-recall curves for pixel-level segmentation. Notice that I employed a RF with only 100 trees, compared to the method proposed by Rigamonti et al.[111] in which 600 trees were used, hence slower at testing time.	107
6.9	Experiments: segmentation. Probability maps computed on images from DRIVE (first row), BF2D (second row) and VC6 (third row). For each row, from left to right, we report original image, result of the best performing baseline (i.e. "CSC, random init."), proposed method's result and ground truth.	108
A.1	(a) mesentery network used to investigate biological factors related to tortuosity, (b) full resolution mesentery network obtained by interpolating (using splines) sampling points along each vessel.	130
A.2	Tortuosity maps obtained by colour encoding the mean curvature at spatial scale 1 (a) and 10 (b).	131
A.3	Density map of the mesentery network in Figure A.1(b).	132

List of Tables

2.1	Curvilinear structure segmentation methods most related to the work reported herein. Asterisk means performance validation was carried out in a different setting compared to other methods.	10
2.2	Tortuosity estimation work most related to the one reported herein. SOAM stands for "sum of angles measure", ICM for "inflection count measure", others are reported in this section.	14
4.1	Effect of patch and dictionary size on the area under precision-recall curves (mean/standard deviation). Our method outperforms the baseline in all conditions.	55
4.2	Parameters of the proposed segmentation architecture.	58
5.1	Tortuosity estimation: comparison of tortuosity estimation algorithms at a parity of segmentation approach. For tortuosity ground truth, each experienced observer is taken as reference in turn (Obs_1 , Obs_2 and Obs_3). The segmentation approach used for these experiments is our " <i>SCIRD, multi-range context</i> " shown to outperform others methods on the IVC140 data set (see Figure 4.12). The first two columns in each table show the performance of the others observers against the one used as reference. The other columns in each table report several tortuosity estimation algorithms. Specifically, <i>MSMW</i> is the algorithm based on multi-scale-multi-window tortuosity [5], <i>MSSPLINE</i> is the proposed approach based on multi-scale rotation invariant spline fitting (results in boldface), <i>MSSPLINE (no FS)</i> is the latter approach without using tortuosity feature selection, TD is the tortuosity density index [54], <i>DM</i> is the distance measure [64], <i>SCC</i> is the tortuosity estimation algorithm based on slope chain coding [23] and τ_5 is the normalised integral of the squared curvature [61].	72

5.2	Tortuosity estimation: comparison of segmentation algorithms at a parity of tortuosity quantification approach. For tortuosity ground truth, each experienced observer is taken as reference in turn (Obs_1 , Obs_2 and Obs_3). The tortuosity quantification approach used for these experiments is our <i>MSSPLINE</i> shown to outperform other methods on our data set (see Table 5.1). The first two columns in each table show the performance of the others observers against the one used as reference. The other columns in each table report tortuosity quantification performance when the following approaches are used: manual segmentation (<i>Manual</i>); the proposed “ <i>SCIRD, multi-range context</i> ” segmentation algorithm (<i>Proposed</i> , in bold face), based on curved-support and multi-range context filters; the segmentation method proposed by [111], based on “locally-straight” and appearance filters.	74
5.3	Confusion matrices for the tortuosity estimation task obtained using the proposed method. For tortuosity ground truth, individual observers are used in turn (Obs_1 , Obs_2 and Obs_3).	75
5.4	Cohen’s kappa and modified kappa for the tortuosity estimation task. For ground truth, individual observers are taken as reference in turn (Obs_1 , Obs_2 and Obs_3). We compare the performance of the proposed method (i.e., <i>SCIRD, multi-range context</i> + <i>MSSPLINE</i> , indicated as <i>Ours</i> here) with the expert observers, in terms of Cohen’s kappa (K , first row) and modified kappa (K / K_M , second row).	75
5.5	Comparison in terms of running time for each tortuosity estimation algorithm alone and in combination with the automated segmentation algorithm proposed herein. The proposed spline-based curvature estimation, i.e. <i>MSSPLINE</i> (results in boldface), is significantly faster than our previous multi-window solution, i.e. <i>MSMW</i> . Experiments were carried out on Intel i7-4770 CPU @ 3.4 GHz, using MATLAB code. Each image was 384×384 pixels. . . .	75
6.1	Total time to convergence (in minutes) for the CSC phase initialised with the proposed method (Prop.), and the baselines. In brackets, the proposed warm start processing time (in seconds).	100
6.2	Comparison in terms of AUPRC, F-measure, Jaccard index and training time (in minutes), between random, DCT-based and the proposed initialisation strategy (denoted as “Ours”).	102

6.3	Influence of the sparsity parameter λ on the segmentation performance (AUPRC) of the proposed initialisation method (<i>Prop.</i>) and the best baseline method (<i>Random</i>) on DRIVE.	103
A.1	Experiment (1): measure the correlation of the best combination of tortuosity features with each haemodynamic parameter separately. "NS" replaces correlations whose p-value was greater than 0.01.	133

Chapter 1

Introduction

This thesis focuses on the automated analysis of curvilinear structures in medical images to improve the quantification and investigation of potential biomarkers, with particular emphasis on tortuosity. Most of the work described herein was done in collaboration with cornea specialists at MEEI, Harvard Medical School, USA; therefore, my main target was corneal nerve fibre analysis, although experiments with retinal blood vessels and neurites were also carried out to validate the proposed algorithms.

This chapter provides the background and the motivation behind the analysis of curvilinear structures in medical images, from both a clinical and technical perspective. Then, I summarise the main contributions and discuss the organisation of this thesis.

1.1 Background and motivation: curvilinear structure analysis in medical images

Curvilinear structures in the human body perform fundamental tasks such as propagating electrochemical stimulation (e.g., dendrites shown in Figure 1.1(a)), transporting blood (e.g., retinal blood vessels shown in Figure 1.1(b)), and maintaining tissue healthy (e.g., corneal nerve fibres shown in Figure 1.1(c)). Shape abnormalities, e.g. stenosis in blood vessels, may considerably alter these important processes and signal major disorders. For

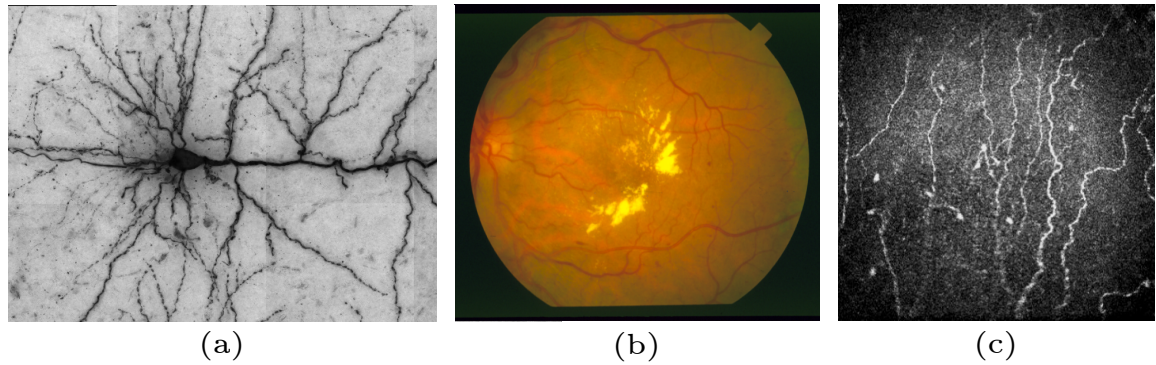


Figure 1.1 Examples of curvilinear structures considered. (a) Neurites; (b) retinal blood vessels; and (c) corneal nerve fibres.

this reason, shape properties of curvilinear structures are carefully assessed in clinical practice to identify *known* abnormalities (diagnosis), determine a therapy and possibly prevent further complications. Moreover, given the key role played by curvilinear structures in the human body and the plethora of non-invasive acquisition modalities currently available, biomarker investigation (i.e., objective indications of medical state measured accurately and reproducibly from outside the patient [127]) based on curvilinear structure shape analysis is emerging as a promising research direction [2, 9, 42, 74, 105, 118, 126, 129, 135, 137].

Some of the most investigated biomarkers for curvilinear structures are width or caliber [118, 137], density [74, 105], fractal dimension [9, 126, 129], bifurcation geometry [2] and tortuosity. In particular, tortuosity has received much attention since the seminal papers by Edington et al. [45] and Cairney [30].

Numerous studies have reported correlations between several pathologies and the tortuosity of a wide range of anatomical structures such as retinal vessels [33, 64, 94, 98, 101, 114, 115], intracerebral vessels [29] and conjunctival blood vessels [104], but also the coronary [47], iliac [38], carotid [19] and aortic [50] arteries, the optic nerve [69] and corneal nerve fibres [46, 57, 58, 72, 79].

The pathologies involved, some of which have high-prevalence, affect a large portion of the population worldwide and include diabetes, diabetic retinopathy and diabetic neuropathy [46, 114, 115], retinopathy of prematurity [64, 136], malignant gliomas [27], facioscapulo-humeral muscular dystrophy [94], spontaneous coronary artery dissection [47], central ret-

inal vein occlusion [101], and children with neurofibromatosis type 1 [69]. In particular, tortuosity has been investigated in corneal diseases such as unilateral herpes zoster [57], herpes simplex keratitis [58], acute acanthamoeba and fungal keratitis [79] and diabetic neuropathy [46, 72].

In several studies (e.g. [47, 57, 81, 101]) tortuosity was assessed by experienced specialists typically grading structures or whole images on a 3-5 level scale or as normal/abnormal based on qualitative, albeit structured, protocols [103]. Regardless of the specific anatomical object of interest, such assessment is subjective and, depending on protocols, tasks and other factors [132], can lead to substantial inter-observer variability and possibly non-negligible intra-observer variability, thus limiting the sensitivity of the assessment scheme. Moreover, requiring direct inspection by specialists limits the amount of images that can be analysed and makes large screening programs unfeasible or at least very expensive.

Several definitions of tortuosity have been proposed to try and quantify tortuosity automatically [28, 43, 54, 61, 72, 131, 136], but no single definition is widely accepted, possibly because tortuosity has different characteristics for different anatomical structures. I argue therefore that a highly adaptable tortuosity estimation algorithm, learning key features for specific image types and structures would be a promising and effective new research target.

However, efficiency would still be limited if manual segmentation were required. This is especially true considering the image resolution of state-of-the-art instruments, e.g. $3,500 \times 2,300$ pixels for standard fundus camera images [26]. Fast and accurate curvilinear structure segmentation is therefore needed, but different characteristics of tortuous curvilinear structures across image modalities make segmentation challenging. In fact, highly tortuous objects violate one of the basic assumptions of most tubular structure detectors, namely *locally straight tubular shape* [49, 60, 83, 123]. Further issues include the presence of non-target structures (clutter), low resolution, noise and non-uniform illumination, depending on the imaging modality considered.

Fully automated tortuosity estimation has been proposed for retinal blood vessels, brain vasculature and corneal nerve fibres [64, 70, 77, 116]. Typically, inaccuracies in the segmentation are the main source of inaccurate automatic tortuosity estimates (e.g. [116]),

therefore segmentation algorithms should be designed with particular care. Moreover, the aforementioned methods are based on mathematical definitions of tortuosity providing *fixed* models of a subjective perception. Such measures combine features like the number of inflection points along the structure, and the length to chord ratio. I argue that such hand-crafted tortuosity definitions limit the accuracy of an automated framework of general value (in terms of agreement with clinical judgement) and could ultimately misrepresent the importance of tortuosity as a biomarker.

1.2 Contributions

The research included herein is funded by the REVAMMAD project, part of the European Union's 7th Framework (FP7) Marie Curie Initial Training Network programme. The network includes 8 universities and 2 companies (main partners), and a number of associate partners (including Tufts Medical Center, USA). The overall aim is to combat some of the EU's most prevalent chronic medical conditions using eye imaging. In this framework, I addressed some *technical* issues related to fully automated tortuosity quantification with application to corneal nerve images. Specifically, I decomposed the task of assigning a tortuosity level to a whole corneal image in two sub-tasks: (1) corneal nerve fibre centrelines detection/segmentation, and (2) image-level tortuosity quantification. I improved the state-of-the-art in both sub-tasks and summarise the main contributions below.

1.2.1 Novel curvilinear structure ridge detectors

Highly fragmented and tortuous structures violate two usual assumptions of hand-crafted filters (henceforth, HCFs), i.e. *continuous* and locally *straight* tubular shapes. Here, I introduce novel HCFs, SCIRD (Scale and Curvature Invariant Ridge¹ Detector) and SCIRD-TS (SCIRD for very thin structure), which are simultaneously rotation, scale, contrast, elongation and, unlike the others, curvature invariant. These new HCFs are shown to outperform

¹In image processing, the term *ridge* refers to a narrow and elongated object within an image.

state-of-the-art ones on corneal nerve fibres (target structures), but also on neurites and retinal blood vessels.

1.2.2 Learning single- and multi-range context filters

Hand-designing filters to capture *inter*-object relationships (i.e. *context* filters) is particularly challenging as the configurations to be considered, especially in the medical domain, can vary significantly. To overcome this issue without increasing the computational cost (e.g. unlike the auto-context solution which learns multiple discriminative models sequentially [133]), I propose to learn a *single* classifier (e.g., Decision Forest) taking as input both *appearance* (estimated by HCFs) and *context* information obtained from learned context filters. To improve context modelling and capture multi-range inter-object relationships I introduce multi-range context filters.

1.2.3 Accelerating convolutional sparse coding for filter learning

Deep learning has shown great potential for curvilinear structure (e.g. retinal blood vessels and neurites) segmentation as demonstrated by a recent auto-context regression architecture based on filter banks learned by convolutional sparse coding [122] (henceforth, CSC). However, learning such filter banks is very time-consuming, thus limiting the amount of filters employed and the adaptation to other data sets (i.e. slow re-training). Driven by the observation that filter banks obtained by CSC applied to curvilinear structures often incorporate filters closely resembling hand-crafted ones, I present a novel approach to accelerate CSC based on refining carefully designed HCFs (warm-start strategy).

1.2.4 Multi-scale analysis of tortuosity measures

To the best of my knowledge, none of the previously proposed tortuosity estimation approaches investigated the role played by the specific spatial scale in tortuosity quantification. I introduce the concept of multi-scale tortuosity representation and show that it is more

suitable for tortuosity estimation, as it takes into account the different contribution of high- and low-frequency turns, often characterising corneal nerve fibres, for instance.

1.2.5 A machine learning approach to tortuosity definition

Several definitions of tortuosity (a.k.a. indices) have been proposed to try and quantify tortuosity automatically, but no single definition is widely accepted, possibly because tortuosity has different characteristics for different anatomical structures. Moreover, these tortuosity indices are fixed a priori and deemed to be suitable for several curvilinear structures and pathologies. I argue that a versatile definition of tortuosity is more suitable to address the tortuosity estimation task. Therefore, I present a machine learning approach to *tortuosity definition*, learning key features for specific image types and structures.

1.3 Thesis organisation

This thesis is organised as shown in Figure 1.2.

- **Related Work.** Chapter 2 reviews work on curvilinear structure analysis in medical images, modelling, segmentation and tortuosity quantification. It also discusses some of the limitations of previous methods which motivated this work.
- **Materials and experimental protocol.** Chapter 3 describes the data sets and protocol used to validate curvilinear structure segmentation approaches and tortuosity estimation. Notice that, although the main focus was corneal nerve fibre analysis, data sets including neurites were used to validate segmentation algorithms, while others including retinal blood vessels were used to validate both segmentation and tortuosity estimation approaches. Moreover, performance evaluation criteria are introduced and discussed, both for segmentation and tortuosity estimation.
- **Curvilinear structure modelling and segmentation.** Chapter 4 presents the novel ridge detector designed for tortuous and fragmented structures (SCIRD), and the approach combining HCFs with learned filters efficiently and effectively (single- and

multi-range context filters). The discussion includes experiments to validate these approaches.

- **Tortuosity estimation.** Chapter 5 discusses a machine learning approach to tortuosity definition and estimation, and a novel representation tool called as tortuosity plane, leveraging tortuosity interpretation, developed in collaboration with clinical partners at Harvard Medical School. A substantial number of experiments validating the whole tortuosity estimation system and the impact of each module on the tortuosity estimation performance are presented and compared with state-of-the-art segmentation and tortuosity estimation algorithms.
- **Improving Curvilinear Structure Modelling and Segmentation.** Chapter 6 explains the reason why the original SCIRD filter banks tend to miss very thin structures and discusses a novel formulation of SCIRD, SCIRD-TS, addressing such limitation. Moreover, this chapter discusses a novel approach to speed-up convolutional sparse coding for filter learning which could be employed to improve detection performance by learning more discriminative context filters.
- **Conclusions, discussion and future work.** Chapter 7 concludes this thesis and suggests future research directions for exploration.
- **Further work carried out during the project.** Appendix A discusses a short project on identifying the biological factors influencing tortuosity, carried out at Charité - Universitätsmedizin Berlin (DE).

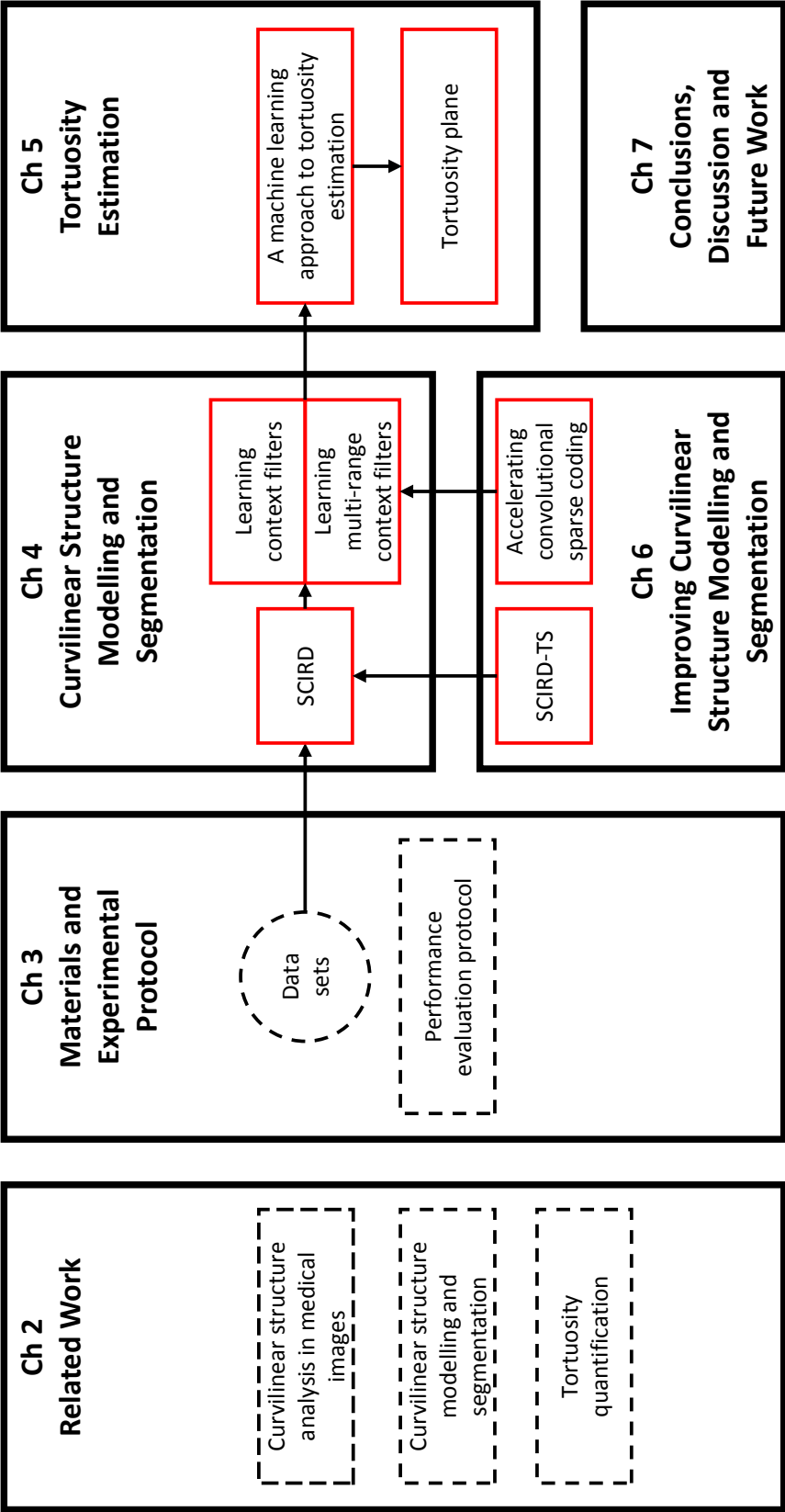


Figure 1.2 Thesis roadmap: chapters and their relation.

Chapter 2

Related Work

2.1 Introduction

This chapter reviews work related to curvilinear structure modelling, segmentation and tortuosity quantification, and discusses the limitations which have motivated this work.

In Section 2.2, I discuss the main challenges behind curvilinear structure segmentation and how these have been addressed so far. I categorise and discuss modelling and segmentation approaches based on the specific feature extraction technique adopted, as most of the methods proposed herein focus on this aspect.

Section 2.3 reviews previous approaches to automated tortuosity quantification. In particular, I present some of the most used tortuosity indices (used as baselines during my experimental validation) and report information about data sets used, analyses carried out and performance measures adopted.

In Section 2.4 I discuss the main limitations of previous modelling, segmentation and tortuosity quantification methods which motivate the proposed solutions.

Section 2.5 summarises this chapter and introduces the next one.

Table 2.1 Curvilinear structure segmentation methods most related to the work reported herein. Asterisk means performance validation was carried out in a different setting compared to other methods.

AUTHORS	DATA	FEATURES	CLASSIFICATION	PERFORMANCE EVALUATION
Frangi et al. [49]	2D angiography (X-ray) 3D MRA	HCFs	Unsupervised	Qualitative
Hoover et al. [66]	STARE (20 images)	HCFs	Unsupervised	ROC, AUROC (0.7590), Acc (0.9275)
Staal et al. [125]	DRIVE (40 images) STARE (20 images)	HCFs	Supervised	ROC, AUROC (DRIVE = 0.9520, STARE = 0.9614), Acc (DRIVE = 0.9516, STARE = 0.9441)
Soares et al. [123]	DRIVE (40 images) STARE (20 images)	HCFs	Supervised	ROC, AUROC (DRIVE = 0.9614, STARE = 0.9671), Acc (DRIVE = 0.9466, STARE = 0.9480)
Mendonca et al. [100]	DRIVE (40 images) STARE (20 images)	HCFs	Unsupervised	Acc (DRIVE = 0.9463, STARE = 0.9479)
Ricci et al. [110]	DRIVE (40 images) STARE (20 images)	HCFs	Supervised	ROC, AUROC (DRIVE = 0.9633, STARE = 0.9680), Acc (DRIVE = 0.9595, STARE = 0.9646)
Law et al. [83]	Synthetic data MRA (1 volume)	HCFs	Unsupervised	Qualitative
Al-Diri et al. [3]	DRIVE (40 images) STARE (20 images)	HCFs	Unsupervised	Se (DRIVE = 0.7282, STARE = 0.7521), Sp (DRIVE = 0.9551, STARE = 0.9681)
Law et al. [84]	Synthetic data, MRA (3 volumes) CTA (1 volume)	HCFs	Unsupervised	Qualitative
Lam et al. [82]	DRIVE (40 images) STARE (20 images)	HCFs	Unsupervised	ROC, AUROC (DRIVE = 0.9614, STARE = 0.9739), Acc (DRIVE = 0.9472, STARE = 0.9567)
Marin et al. [96]	DRIVE (40 images) STARE (20 images)	HCFs	Supervised	ROC, AUROC (DRIVE = 0.9588, STARE = 0.9769), Acc (DRIVE = 0.9452, STARE = 0.9526)
Rigamonti et al. [111]	DRIVE (40 images) STARE (20 images) BF2D (2 images) VC6 (3 images)	HM	Supervised	PRC
Ganin et al. [51]	DRIVE (40 images)	DLA	Supervised	PRC, AUPRC (DRIVE = 0.89)
Annunziata et al. [4]	STARE (20 images) HRF (45 images)	HCFs	Unsupervised	AUROC (STARE = 0.9655), Acc (STARE = 0.9562, HRF = 0.9581)
Sironi et al. [121]	DRIVE (40 images) BF2D (2 images) OPF (volumes)	DLAs	Supervised	AUROC (DRIVE = 0.962, BF2D = 0.98, OPF = 0.997), F-measure (DRIVE = 0.786, BF2D = 0.749, OPF = 0.567)
Sironi et al. [122]	Brightfield (5 volumes) VC6 (5 volumes) Vivo2P (5 volumes)	DLA	Supervised	PRC (pixel-based evaluation), others (tracing evaluation)
Sironi et al. [120]	DRIVE (40 images)	DLA	Supervised	PRC, F-measure (0.81)
Gu et al. [55]	DRIVE (40 images) STARE (20 images) Neuronal (112 images)	DLA	Supervised	Acc (DRIVE = 0.9732*, STARE = 0.9772*), modified F-measure (STARE = 0.8092, Neuronal = 0.868)
Zhao et al. [138]	DRIVE (40 images) STARE (20 images) VAMPIRE (8 images)	HCFs	Unsupervised	AUROC (DRIVE = 0.862, STARE = 0.874, VAMPIRE = 0.857), Acc (DRIVE = 0.954, STARE = 0.956, VAMPIRE = 0.977)
Azzopardi et al. [10]	DRIVE (40 images) STARE (20 images) CHASEDB1 (28 images)	HCFs	Unsupervised	ROC, AUROC (DRIVE = 0.9614, STARE = 0.9563, CHASEDB1 = 0.9487), Acc (DRIVE = 0.9442, STARE = 0.9497, CHASEDB1 = 0.9387)
Li et al. [89]	DRIVE (40 images) STARE (20 images) CHASEDB1 (28 images)	DLA	Supervised	ROC, AUROC (DRIVE = 0.9738, STARE = 0.9879, CHASEDB1 = 0.9716), Acc (DRIVE = 0.9527, STARE = 0.9628, CHASEDB1 = 0.9581)

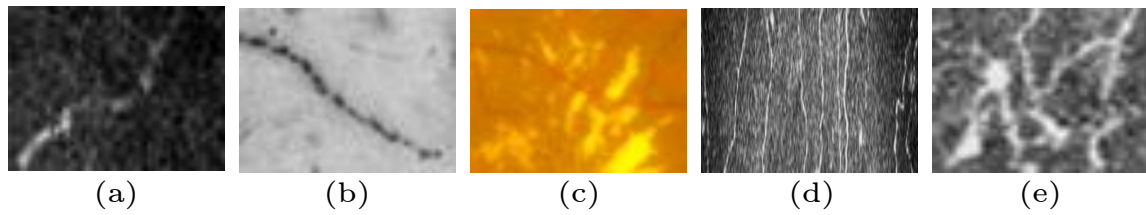


Figure 2.1 Factors making curvilinear structure segmentation challenging. (a) a corneal nerve fibre with low signal-to-noise ratio; (b) a fragmented neurite; (c) retinal blood vessels surrounded by exudates (i.e., confounding non-target structures); (d) corneal nerve image with non-uniform illumination; and (e) corneal nerve fibres with complex configurations (e.g., tortuosity, bifurcations and parallel fibres).

2.2 Curvilinear structure modelling and segmentation

Curvilinear structure segmentation is a particularly active area of research [3, 4, 6, 7, 49, 51, 55, 66, 82–84, 89, 96, 100, 110, 111, 120–123, 125, 138]. The problem is made challenging by multiple factors, as shown in Figure 2.1, including low signal-to-noise ratio at small scales, variable appearance, confounding non-target structures, non-uniform illumination and complex configurations. In an attempt to address such challenges, several segmentation methods have been proposed (see [88] for an extensive review). Table 2.1 summarises the curvilinear structure segmentation methods most related to the work reported herein. A common way to categorise such methods is based on the type of classifier used to obtain pixel-level predictions, i.e. unsupervised or supervised (see e.g. [10]). The latter methods tend to outperform the former ones in terms of detection performance, but they are often slower due to the classifier (e.g. neural network or decision forest). Moreover, training a classifier (supervised methods) often requires a substantial amount of labelled training data (i.e. manually segmented images) which may be difficult to procure. For these reasons, adopting one solution or the other depends on the application itself and the “speed/detection performance” trade-off.

The extraction of adequate features is a fundamental aspect for segmentation approaches and, since most of the methods proposed herein focus on this aspect, I categorise them based on the feature extraction approach adopted, namely hand-crafted filters, deep learning architectures and hybrid approaches.

2.2.1 Hand-crafted filters

Hand-crafted filters (HCFs) model local geometrical properties of ideal tubular structures. Eigenvalue decomposition of the Hessian matrix was employed by [4, 49, 97, 113]; maximum projections over each scale were used to make the approach scale-invariant. These projections were then used to build the well-known tubularity measure called *vesselness* by Frangi et al. [49]. However, performance tends to degrade at crossings or bifurcations since this approach only looks for elongated structures. To overcome this issue, Hannink et al. [60] proposed to segment crossings/bifurcations with multiscale invertible orientation scores and apply vesselness filters to maps of the latter. Optimally Oriented Flux (OOF) was recently proposed by Law and Chung [83] to improve detection of adjacent structures with vesselness measures. OOF is based on the computation of an optimal projection direction minimizing the inward oriented flux at the boundary of localized circles (2-D) or spheres (3-D) of different radii. This projected flux can be regarded as a likelihood that a pixel is part of a tubular structure. Tubularity measures can be obtained by combining the eigenvalues of the OOF Hessian matrix. Other successful HCFs rely on Gabor wavelets (a.k.a. Morlet wavelets) as proposed, for instance, by Soares et al. [123] who exploited their directional selectiveness to detect oriented structures and achieve fine tuning to specific frequencies. HCFs have also been used by Honnorat et al. [65] to compute a local tubularity measure feeding a graphical model.

2.2.2 Fully learned architectures

Fully learned architectures (FLAs) [16, 17, 86, 117] have shown the potential to overcome HCFs' modelling issues by learning object representations directly from training data, with excellent performance reported on several tasks [48, 78]. A key difference with respect to traditional approaches is that the intermediate features/filters and classifiers are learned jointly, with the aim of maximising segmentation performance. This results in a large quantity of parameters (and many hyper-parameters) to be optimised; therefore, training such complex learning architectures from scratch requires high-performance hardware and/or op-

timised implementations, and more importantly large datasets to avoid over-fitting, which are not always available for medical images. For this reason, Becker et al. [15] have recently proposed a less complex solution employing gradient boosting to learn convolutional filters and boosting weights simultaneously, applied with success to retinal blood vessels and neurites. Sironi et al. [119] have used the responses of convolutional filters learned by sparse coding as input features to multiple regressors trained to predict the distance from the centreline. Recently, Li et al. [89] successfully used convolutional neural networks (CNN) for retinal vessel segmentation and showed improved performance over many state-of-the-art methods based on traditional (HCFs + classifier) approaches. Training time and over-fitting are limited by unsupervised pre-training (using an auto-encoder). The initial result (i.e. filters in the first layer and all the weights of the CNN) is then refined using standard CNN optimisation (i.e. backpropagation).

2.2.3 Hybrid methods

Hybrid methods (HMs) combine HCFs with filters learned by FLAs, exploiting the efficiency of fast HCFs while limiting the amount of learned filters. The first HM applied to tubular structures was proposed by Rigamonti et al. [111] and it is based on feature vectors obtained by concatenating OOF filter responses (i.e., HCFs) with those of learned *appearance* filters, i.e. learned on the original training images. It employs convolutional sparse coding (CSC) to learn 9 appearance filters. Quantitative results show a clear improvement over methods based only on HCFs, achieving the same level of performance obtained with a filter bank of 121 filters learned via SC, at a limited computational cost. Notice that several days are reportedly needed to learn 121 filters by CSC (without parallel or GPU processing) [111], while 9 filters are learned in less than 30 minutes thus making adaptation to other data sets (re-training) fast without worsening segmentation performance.

2.3 Tortuosity quantification

Automated tortuosity quantification is particularly challenging due to the variability of anatomical tubular structures and to the varying importance of each tortuosity feature (e.g., curvature, number of inflection points) for different structures of interest (e.g., retinal vessels, coronary artery) within specific clinical contexts (e.g., diabetes, malignant gliomas, facioscapulohumeral muscular dystrophy). Most of the algorithms reported here focus on the analysis of curvilinear structure centrelines, as the role played by the caliber/width is still not clear [131]. Table 2.2 summarises the semi-automated and automated tortuosity quantification systems most related to the work reported herein, along with specific data,

Table 2.2 Tortuosity estimation work most related to the one reported herein. SOAM stands for "sum of angles measure", ICM for "inflection count measure", others are reported in this section.

AUTHORS	DATA	FEATURES	ANALYSIS	PERFORMANCE MEASURE
Heneghan et al. [64]	Retina (23 subjects)	DM (and width)	retrospective normal/abnormal (subject-level)	ROC, Se (0.82), Sp (0.75)
Bullitt et al. [28]	Brain (18 volumes)	SOAM ICM	retrospective normal/abnormal (subject-level)	Mean, SD
Grisan et al. [54]	Retina (60 vessels)	TD	rank correlation	Spearman (arteries = 0.949, veins = 0.853)
Hart et al. [61]	Retina (20 images)	curvature-based (τ_5)	2-class classification (vessel- and network-level)	ROC (vessels = 0.96, networks = 0.86), % correctly class. (vessels = 89.5, networks = 90)
Wilson et al. [136]	Retina (75 vessels)	Incremental length-based	rank correlation	Spearman (0.673)
Trucco et al. [131]	Retina (200 vessels)	Curvature-caliber	3-class classification	confusion matrices
Turior et al. [134]	Retina (60 images)	Chain coding	3-class classification	Se (86.7), Sp (96.7), Acc (91.4), PPV (92.9), NPV (93.7)
Joshi et al. [70]	Retina (8 patients)	TD-based	rank correlation (4-class grading)	Spearman (arteries = 1, veins = 0.77)
Poletti et al. [109]	Retina (20 images)	Linear comb. of indices	rank correlation	Spearman (0.95)
Ghadiri et al. [52]	Retina (10 images), conjunctiva (10 images)	Curvature-based (filtering)	rank correlation	Spearman (Retina = 0.94, conjunctiva = 0.89)
Kallinikos et al. [72]	Cornea (36 images)	TC	retrospective	Mean, SD
Scarpa et al. [116]	Cornea (30 images)	TD-based	3-class classification	Krippendorff (0.96), % correctly class. (0.93)

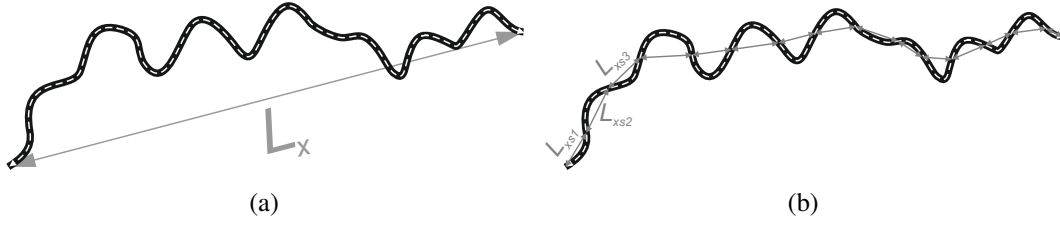


Figure 2.2 (a) Key quantities for the DM; (b) vessel partitioning used by the TD algorithm.

tortuosity index/definition (features), type of analysis and performance measure [28, 52, 54, 61, 64, 70, 109, 131, 134, 136] (refer to [71] for an extensive review). Here I report several tortuosity indices/definitions which will be used as baselines in Chapter 5; indices were selected considering together their performance (as reported in the original papers), recency and impact on the field (citations, take-up)[93].

Distance Measure (DM). The DM is one of the earliest tortuosity indices [61, 64]. Its popularity may depend on its simple and intuitive definition:

$$DM = \frac{L_c}{L_x} \quad (2.1)$$

where L_c is the vessel centreline length (dashed white line in Figure 2.2(a)) and L_x the chord length (line joining the vessel's endpoints, Figure 2.2(a)). DM is 1 when a vessel is perfectly straight and increases with tortuosity. The limits of the DM have been pointed out previously [31, 61], crucially its inability to distinguish vessels with multiple bends (very tortuous) from vessels with a single arc (less tortuous) that have the same average deviation from the chord. This problem arises because DM is a global index which fails to capture local changes.

Tortuosity Density (TD). To address the problem above, Grisan et al. proposed the TD index [54] (Figure 2.2(b)). TD assesses vessel tortuosity by summing the contributions to tortuosity of uniformly convex or concave arcs, as follows:

$$TD = \frac{n-1}{n} \frac{1}{L_c} \sum_{i=1}^n \left[\frac{L_{csi}}{L_{xsi}} - 1 \right] \quad (2.2)$$

Here, n is the number of “turns” (curvature sign changes, i.e., zero-crossings of the second derivative of the centreline), L_{csi} is the arc length of segment i , L_{xsi} is the chord length of segment i . L_c is the length of the whole vessel centreline. A vessel centreline with only one turn has $TD = 0$; with more than one turn, the tortuosity is greater than 0 (avoiding the problem of DM). The TD index is also normalized to vessel length ($1/L_c$), which allows comparison of vessels of various lengths and scale invariance. Among several curvature-based and inflection point-based tortuosity indices, the authors found TD to be the most accurate index to model clinical scores of retinal vessel tortuosity with hypertensive retinopathy images [54].

Slope Chain Coding (SCC). The SCC index was introduced recently as a general index for planar curves by Bribiesca [23]. The original paper includes a qualitative demonstration on a single retinopathy of prematurity (ROP) image, but a more comprehensive experimental validation was carried out in [93]. To calculate SCC, a vessel centreline is approximated by a linear piecewise curve formed by line segments of fixed length, and the slope change between segments is computed. The SCC index is defined as the sum of the absolute values of the slope changes along the centreline:

$$SCC = \sum_{i=1}^n |a_i|. \quad (2.3)$$

Here, n is the number of slope changes (the number of segments minus one) and a_i is the slope change after the i -th segment. The influence of n on the tortuosity is assessed in [23] with an example showing invariance of SCC with two values of n , and subsequently investigated in [93] on retinal vessels graded by experts.

Curvature-Integral Measures. Curvature-based measures were introduced by Hart et al. [61]. Here I report the best-performing one ([93]), defined as

$$\tau_5 = \frac{1}{L_c} \int_{t_0}^{t_1} k^2(t) dt \quad (2.4)$$

where L_c is the vessel centreline length, t is the curvilinear coordinate and $k(t)$ is the curvature defined as

$$k(t) = \frac{x'(t)y''(t) - x''(t)y'(t)}{[y'(t)^2 + x'(t)^2]^{3/2}}. \quad (2.5)$$

2.3.1 Corneal nerve fibres tortuosity assessment

Several quantitative measures of the vascular tortuosity have been proposed. Comparatively little work exists on quantitative tortuosity measures for corneal fibres. Kallinikos et al. [72] were the first to propose an objective, semi-automated method for quantifying sub-basal nerve tortuosity. The first and second derivatives of the function representing fibre centrelines are squared and added. The sum is multiplied by the length of the interval (x_j, x_{j+1}) , to estimate the change in the direction of the nerve fibre, within that interval. The sum of all the values is obtained and the square root taken. Once all the quantities have been computed, the tortuosity TC is calculated by the following formula:

$$TC = \sqrt{\sum_{j=1}^{n-1} (x_{j+1} - x_j) \{ [f'(x_j)]^2 + [f''(x_j)]^2 \}} \quad (2.6)$$

where $f'(x_j)$ and $f''(x_j)$ are the first and second derivatives at the point x_j , respectively.

Scarpa et al. [116] adopted Grisan et al.'s algorithm [54] and reported an experimental comparison of tortuosity measures; their results suggest that the proposed algorithm was the one with best associations with annotations provided by a cornea specialist.

2.4 Discussion

In this chapter I briefly reviewed the work related to curvilinear structure modelling, segmentation and tortuosity estimation. It is worth noting that most of the curvilinear structure segmentation approaches reported so far have not yet been explored for corneal nerve fibres, my target structures. Moreover, most tortuosity estimation algorithms have been proposed and validated on blood vessels, with only a few exceptions ([72, 116]) adjusted to work

on corneal nerve fibres. Below I discuss the main limitations of existing approaches for modelling, segmentation and tortuosity estimation which motivates the proposed solutions.

2.4.1 Limitations: Modelling and Segmentation

Modelling appearance properties of curvilinear structures inevitably requires making assumptions that might be violated in some cases. For instance, highly fragmented and tortuous structures violate two usual assumptions of most HCF models, i.e. continuous and locally straight tubular shapes. While discontinuity can be addressed by adopting elongated filters (e.g., Gabor [123]), no *hand-crafted* ridge detector for locally non-straight tubular shapes has been proposed so far. Given my target application, i.e. tortuosity estimation (hence tortuous structures), this implicit assumption is expected to have a negative impact on segmentation performance and therefore requires particular care when designing the segmentation pipeline. This is confirmed by my comparative experiments (Chapter 4).

Although FLAs are designed to overcome these modelling limitations, they tend to require a large amount of training data, including manual annotations, which is not always available in the medical domain. In this setting, HMs seem to be more suitable to address the segmentation task. However, the solution proposed by Rigamonti et al. [111] learns filters independent of the HCFs used, so that some filters may resemble the ones already included in the HCF banks. This would result in a subset of redundant filters which may reduce the discriminative power of the entire feature set and affect segmentation performance. Another limitation of this HM is that context information (i.e. inter-object relationships) is not taken into account, while it has been shown to have a significant positive effect on segmentation performance [133]. However, the auto-context method proposed by Tu and Bai [133] typically requires learning multiple discriminative models (i.e. classifiers) sequentially, thus leading to slower predictions at test time. Recently, an auto-context framework (multi-layer) based on unsupervised filter learning has been shown to outperform CNN and modifications [51] on curvilinear structure segmentation in the medical domain [120, 122]. The framework proposed in [120, 122] relies on filters learned through CSC [111, 121], but learning them is very time-consuming [111]. Therefore, the filter bank learned at the

first layer is kept unchanged across the other ones, due to the prohibitive cost of learning layer-specific filter banks [122]. This represents the first limitation as learning layer-specific filters would model higher-order properties of curvilinear structures and potentially improve segmentation performance. Second, since the visual appearance of curvilinear structures may vary significantly and the range of acquisition modalities may lead to different image characteristics in terms of contrast and noise, re-training (learning new filter banks) is often necessary to achieve good performance. Given the time required to learn such filter banks, re-training would be relatively slow.

Finally, it is worth noting that there is no widely accepted validation protocol for such methods (Table 2.1). Although accuracy (Acc), ROC and AUROC have been used in early work, several authors have pointed out that these measures are not suitable to tease out differences among methods (Acc>0.9 can be achieved by simply predicting always “background”) due to imbalance of negative and positive instances (i.e. number of background and vessel/fibre pixels, respectively). To overcome these issues, precision-recall curves, F-measure and AUPRC are currently the measures of choice to assess segmentation performance.

2.4.2 Limitations: Tortuosity estimation

The main limitation of tortuosity estimation approaches lies in the fact that, although several attempts have been made to try and define tortuosity, no widely accepted definition is currently available. I argue that this is due to the variability of tortuosity characteristics among structures and, possibly, pathologies. Therefore, the definitions of tortuosity proposed previously, which are defined a priori, do not seem to be suitable to capture such variations. Some of the best-performing tortuosity indices are based on a combination of a set of tortuosity measures, such as DM and number of inflection points. The relative weight of each tortuosity measure in this combination is again fixed a priori and deemed to be suitable in general for different structures and pathologies.

An in-depth visual investigation of curvilinear structures such as corneal nerve fibres suggests that multiple frequency components contribute to tortuosity. Previously proposed

approaches ignore this aspect and compute tortuosity features as if only a single frequency component was present.

Curvature estimation algorithms adopted by the best-performing tortuosity indices are typically based on finite differences leading to noisy estimates, especially when segmentations are obtained automatically. Methods based on chain coding such as SCC do not seem to solve the problem, given reported experimental results.

Based on the information reported in Table 2.2, the test set (for image- or patient-level tortuosity estimation) is often rather small, including at most 60 retinal images and 30 corneal nerve images (36 images for a semi-automated algorithm). Moreover, when images are graded with different *levels*, no more than 3 classes are used, with the exception of [70], assessing the performance in terms of rank correlation for only 8 images split in 4 levels.

2.5 Conclusions

In this chapter, I have discussed the main approaches for curvilinear structure modelling, segmentation and tortuosity quantification.

Methods for curvilinear structure modelling and segmentation have been categorised based on the feature extraction technique adopted, as most of the methods proposed herein focus on this aspect. Traditional segmentation approaches are mostly based on carefully designed HCFs. More recently, FLAs have been proposed and shown to overcome some of the HCF-based methods' limitations. However, FLAs require a much higher computation power to achieve state-of-the-art performance and, more importantly, may require large amounts of annotated data. HMs are emerging as a potential solution to mitigate such requirements (especially the one related to the need for annotated data).

The main difficulty of tortuosity quantification is the lack of a widely-accepted definition. This has led to the proposal of several tortuosity indices some of which, have been reviewed in this chapter. Most of these indices have been adopted for retinal vessel tortuos-

ity estimation, but also for brain and conjunctival vessels. Little work has been reported in the literature on corneal nerve fibre tortuosity quantification, my main target structure.

In the next chapter, I will discuss the data sets and the experimental protocols used to validate the proposed fully automated tortuosity estimation method.

Chapter 3

Materials and Experimental Protocol

3.1 Introduction

I used 6 data sets to assess the performance of the various segmentation and tortuosity estimation algorithms discussed in this thesis. In this chapter I briefly describe these data sets and the technical challenges when using them to validate segmentation and/or tortuosity estimation algorithms. Then, I discuss the adopted performance evaluation protocol for each task.

3.2 Data sets

3.2.1 IVCN100

An initial set of 100 2-D images with 384×384 pixels were selected by the clinical collaborators (MEEI, Harvard Medical School) from an existing database of images of the sub-basal nerve plexus in the central cornea, acquired with laser scanning *in vivo* confocal microscopy (IVCM) (Heidelberg Retina Tomograph 3 with the Rostock Cornea Module, Heidelberg Engineering GmbH, Heidelberg, Germany). The diode laser source of this microscope has a 670 nm red wavelength and the microscope is equipped with a $63\times$ objective lens with a numerical aperture of 0.9 (Olympus, Tokyo, Japan). The images obtained by

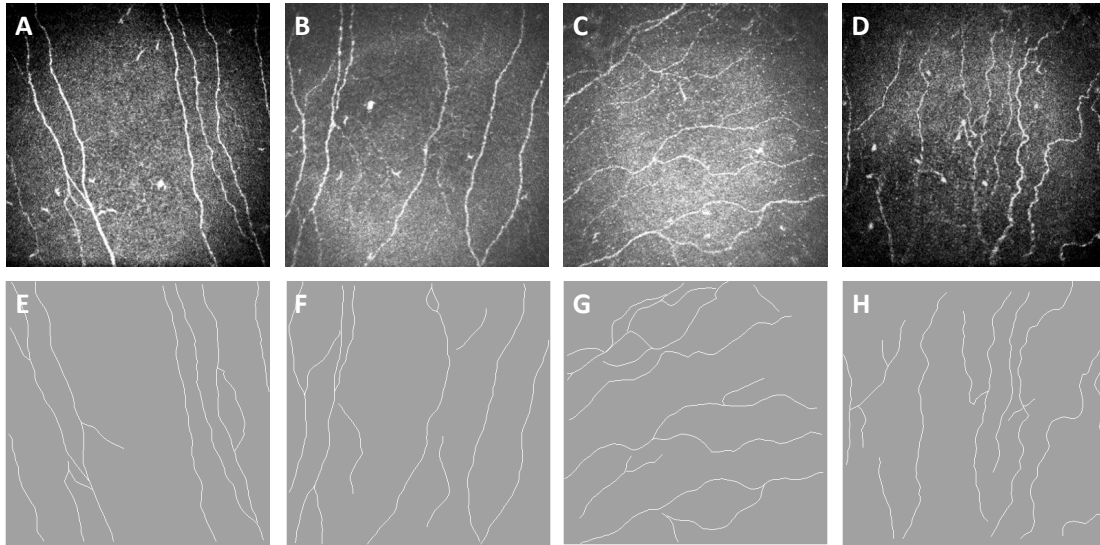


Figure 3.1 Four examples of corneal nerve fibre images captured through *in vivo* confocal microscopy. A,B,C,D: original images with increasing tortuosity level (A = 1, D = 4). E, F, G, H: manually traced fibre centrelines for image A, B, C, D, respectively.

this confocal microscope represent a coronal section of the cornea of $400 \times 400 \mu m^2$ which can be of any corneal layer.

For this general, albeit pilot-level study, images were selected from subjects with Dry Eye Disease (DED) showing a wide spectrum of tortuosity characteristics (Figure 3.1).

For ground truth, corneal nerve fibre centrelines were manually traced by a clinical collaborator using NeuronJ¹, an add-on plug-in for the ImageJ software² [41]. The tortuosity level of each corneal nerve image was determined using a clinically accepted grading scale (grades 1 through 4) [103] by three experienced observers (Dr Ahmad Kheirkhah, Dr Shruti Aggarwal and Dr Pedram Hamrah, clinical authors in [5]) independently.

Segmenting corneal nerve fibres in this data set is particularly challenging due to the presence of poorly contrasted, fragmented and highly tortuous fibres. Moreover, various images contain confounding, non-target structures such as dendritic cells easily mistaken for fibres given similar appearance. Finally, it is worth noting that using this data set for validating segmentation and, more importantly, tortuosity estimation algorithms represents a more challenging scenario compared to previous work. As was shown in Table 2.2, this

¹<http://www.imagescience.org/meijering/software/neuronj>.

²<http://imagej.nih.gov/ij>.

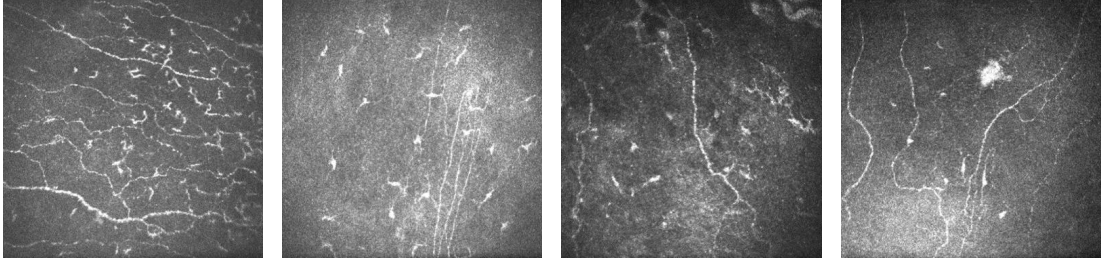


Figure 3.2 Four examples of corneal nerve fibre images from subjects with Herpes Simplex Virus keratitis.

data set is more than 3 times bigger than the one used by Scarpa et al. [116] to validate a fully automated tortuosity estimation method. Moreover, the clinical ordering in 4 tortuosity levels calls for higher discrimination capability compared to the previous fully automated solutions which were tested on 2 or 3 levels (e.g., [61, 116, 131]).

3.2.2 IVCN140

Although a wide spectrum of tortuosity characteristics is shown in IVCN100, it includes only subjects with DED. To make validation more robust, 40 additional images were included to build IVCN140. Specifically, 20 images were taken from healthy subjects and 20 from patients with Herpes Simplex Virus (HSV) keratitis. Detecting corneal nerve fibres in HSV images is particularly challenging as the presence of non-target structures becomes more pronounced and fibres appear thinner and even less contrasted (Figure 3.2).

3.2.3 BF2D

The BF2D dataset [111] consists of two minimum intensity projections of bright-field micrographs that capture neurons (Figure 3.3). The images have a high resolution (1024×1792 and 768×1792) [111], considering the instruments adopted for this kind of imaging, but a low signal-to-noise ratio because of irregularities in the staining process; the dendrites often appear as point-like (fragmented) structures easily mistaken for noise. This data set includes masks to eliminate the nucleus. I adopted the same set partition used in [111],

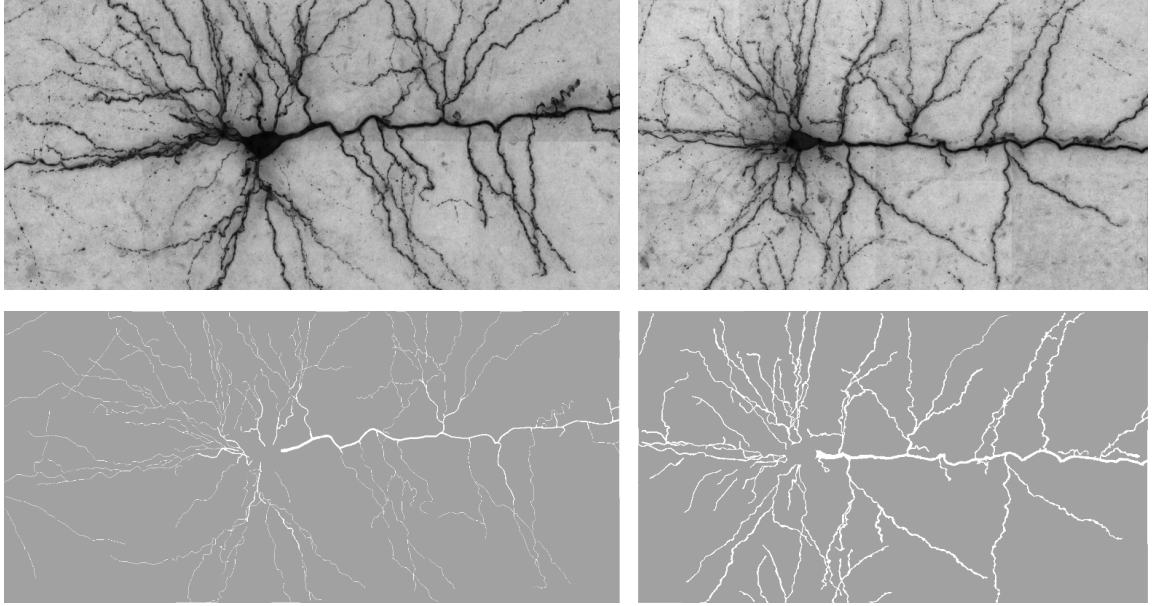


Figure 3.3 Original images (first row) and ground truth segmentation (second row) from the BF2D data set.

retaining one image for training and the other for testing. As was shown in Table 2.1, BF2D is a relatively new data set which has been used as benchmark in [15, 111, 119, 121, 122].

3.2.4 VC6

The VC6 dataset was created by Rigamonti et al. [111] from a set of 3D images showing dendritic and axonal subtrees from one neuron in the primary visual cortex. The original 3D images are part of the publicly available data set used recently for the international DIADEM segmentation challenge (Visual Cortical Layer 6 Neuron) [25]. This data set includes three high-resolution images (882×378 , 630×441 and 817×588 pixels)[111], considering the instruments adopted for this kind of imaging, obtained by computing minimum intensity projections of three image stacks (3-D images), hence showing numerous artefacts, poor contrast and blob-like confounding structures as shown in Figure 3.4. I retained two images for training and the third one for testing, adopting the same set partition used in [111].

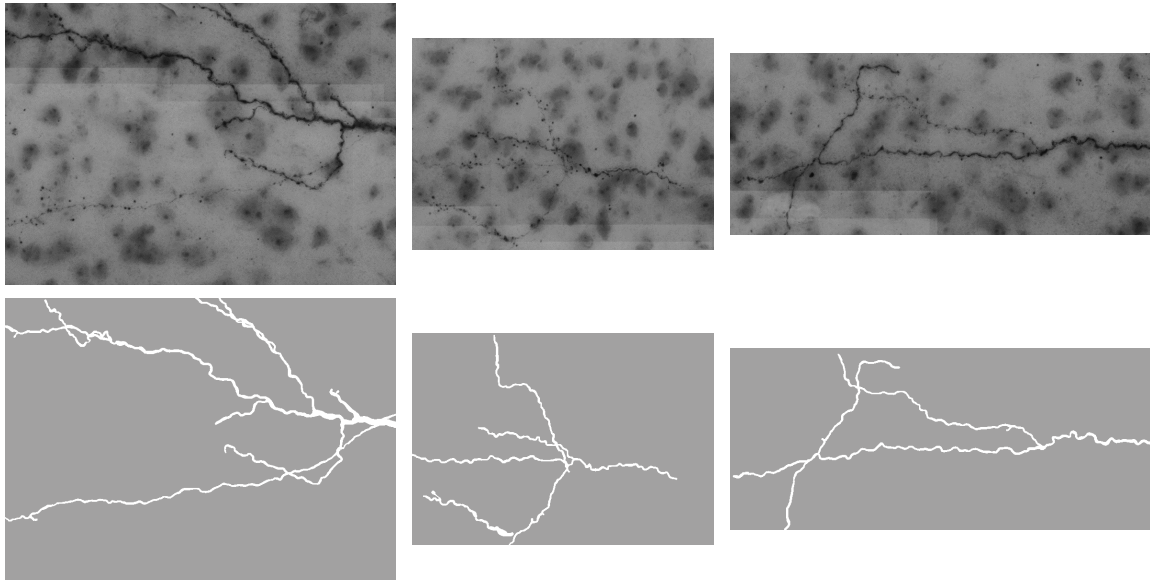


Figure 3.4 Original images (first row) and ground truth segmentation (second row) from the VC6 data set.

3.2.5 DRIVE

DRIVE [125] has been widely adopted as a benchmark data set for vessel segmentation [3, 51, 55, 82, 89, 96, 100, 110, 111, 120, 121, 123, 138]. It includes 40 colour retinal images from a diabetic retinopathy screening program in the Netherlands. The images were acquired by a fundus camera (CR5 non-mydratic 3-CCD, Canon, Tokyo, Japan) with 45 degrees field of view. Images are low-resolution (768×584 pixels), considering the resolution of contemporary fundus camera, hence challenging to segment. The data set was split by Staal et al. [125] into training and testing set, each including 20 images. As shown in Figure 3.5, low resolution, non-uniform illumination, low contrast, vessel central reflex, and confounding structures (exudates, haemorrhages, optic disk) make this data set difficult to segment automatically.

3.2.6 STARE

STARE [66] is another data set including fundus images, widely used as benchmark for retinal vessel segmentation [55, 82, 89, 96, 100, 110, 111, 123, 138]. The full data set includes 397 colour images captured by a TopCon TRV-50 fundus camera at 35 degrees

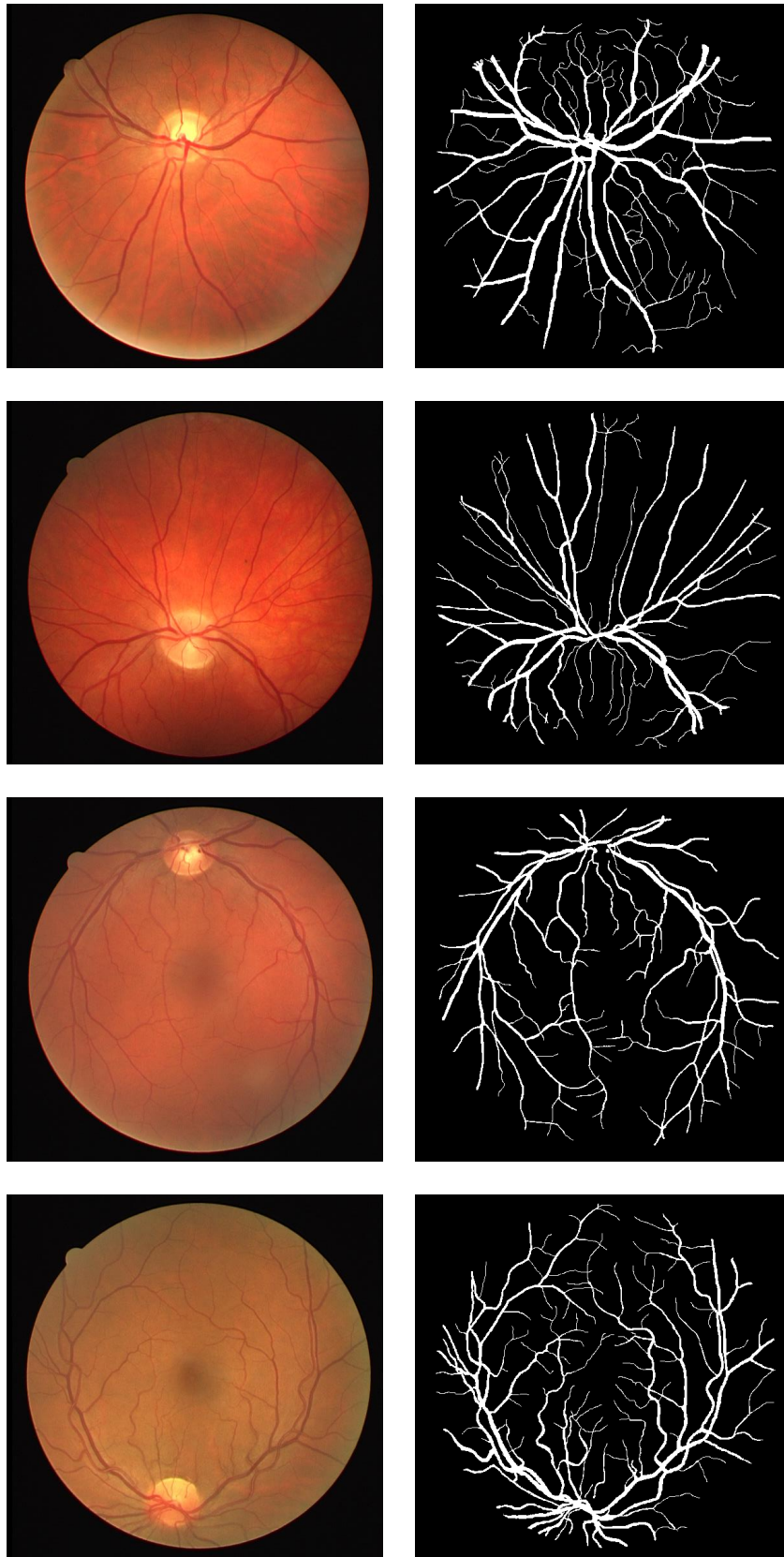


Figure 3.5 Original images (first row) and ground truth segmentation (second row) from the DRIVE data set.



Figure 3.6 Original images (first row) and ground truth segmentation (second row) from the STARE data set.

field of view. Each image is 605×700 pixels. A subset of 20 images (10 normal and 10 abnormal) were manually segmented by two experts [66]. Following the literature (e.g. [96, 123]), I adopted the first observer as ground truth. As shown in Figure 3.6, in addition to the challenges presented by DRIVE, the STARE data set includes images with large abnormal regions and/or a large amount of small abnormal ones, which make this data set difficult to segment automatically.

3.3 Performance evaluation protocol

3.3.1 Segmentation

As discussed in Section 2.4.1, accuracy, ROC and AUROC have been used in early work. However, several authors have pointed out that these measures are not suitable to tease out differences among methods for curvilinear structure segmentation (accuracy greater than 0.9 can be achieved by simply predicting always “background”) due to the imbalance of negative and positive instances (i.e. number of background and vessel/fibre pixels, respectively). To overcome these issues, precision-recall curves (PRCs), AUPRC and F-measure are currently used to assess segmentation performance (e.g., [111, 121, 122]). I adopt the same performance evaluation protocol of the current literature. Specifically, precision and recall are defined as

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.2)$$

where TP indicates true positives, FP false positives and FN false negatives. A curve can be built by measuring precision and recall after applying different thresholds on the output variable. The AUPRC is measured here using the lower trapezoid point estimator [20] and is expressed by a single number summarising the information in the PRC. The F-measure

(a.k.a. F1 score) is defined as

$$\text{F-measure} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (3.3)$$

For centreline detection (used for tortuosity estimation) I adopt a tolerance factor ρ , as done, for instance, by Sironi et al. in [119]: a predicted centreline point is considered a true positive if at most ρ pixels away from the closest ground truth centreline point. All the ground truth centreline pixels that have no predicted centreline pixels within ρ pixels are considered false negatives.

Following the established benchmarking protocol [111, 119], I average performance measures over multiple random sub-sampling cross-validation runs (details are given in the sections reporting the experiments).

3.3.2 Tortuosity estimation

Tortuosity estimation requires assigning images or vessels to a tortuosity level on a given scale, for instance, 1 (normal) to 4 (severe tortuosity). This can be cast as a classification problem. Following the standard performance assessment protocol for multi-class classification [124], we use *weighted* accuracy (Acc), sensitivity (Se) and specificity (Sp), positive predicted value (Ppv) and negative predictive value (Npv), defined below.

$$\text{Acc} = \sum_{i=1}^{N_c} w_i \frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{FN}_i + \text{FP}_i + \text{TN}_i} \quad (3.4)$$

$$\text{Se} = \sum_{i=1}^{N_c} w_i \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \quad (3.5)$$

$$\text{Sp} = \sum_{i=1}^{N_c} w_i \frac{\text{TN}_i}{\text{TN}_i + \text{FP}_i} \quad (3.6)$$

$$\text{Ppv} = \sum_{i=1}^{N_c} w_i \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \quad (3.7)$$

$$\text{Npv} = \sum_{i=1}^{N_c} w_i \frac{\text{TN}_i}{\text{TN}_i + \text{FN}_i} \quad (3.8)$$

where TP_i indicates true positives, TN_i true negatives, FP_i false positives, FN_i false negatives for the i -th tortuosity level; w_i represents the percentage of images whose GT level is i , according to observer taken as reference, as suggested by [124] (N_c , the number of classes/tortuosity levels is 4). The above performance measures do not take into account that the tortuosity level to be predicted is indeed an ordinal variable (i.e. the tortuosity levels are sorted, not independent from each other). In fact, mis-predicting level 4 when the GT tortuosity level is 1, represents a “larger” error than mis-predicting level 2. To quantify this kind of error, I also include the mean squared error (MSE) and mean absolute error (MAE) defined as

$$\text{MSE} = \frac{1}{N_{im}} \sum_{i=1}^{N_{im}} (y_{c_i} - c_i)^2 \quad (3.9)$$

$$\text{MAE} = \frac{1}{N_{im}} \sum_{i=1}^{N_{im}} |y_{c_i} - c_i| \quad (3.10)$$

where N_{im} is the number of test images, y_{c_i} is the predicted tortuosity level, and c_i is the true one for the i -th image.

Since IVC100 and IVC140 data sets (used to validate image-level tortuosity estimation algorithms) were annotated by three experienced observers independently, we take the classification by one observer in turn as ground truth, and compare the performance of ours and baseline methods with that of the other two observers.

3.4 Conclusions

In this chapter, I have described the data sets and evaluation protocol used to validate the proposed segmentation and tortuosity estimation approaches.

The data sets include images of corneal nerve fibres (target structure), but also neurites and retinal blood vessels to test generality. The data sets including corneal nerve fibres

(IVCM100 and IVCM140) have been provided by our clinical collaborators at the Harvard Medical School and they are not public. The data sets including retinal blood vessels (DRIVE and STARE) are public and have been widely adopted as benchmarks for retinal vessel segmentation methods. The data sets including neurites (BF2D and VC6) are emerging benchmark data sets and have been provided by the CVlab-EPFL upon request.

In the next chapter, I will discuss the proposed approach to curvilinear structure modeling and segmentation.

Chapter 4

Curvilinear Structure Modelling and Segmentation

4.1 Introduction

Appearance features can model object-specific shape properties of curvilinear structures and can be hand-crafted, or, more generally, learned directly from a training set of images. Such appearance features are used to enhance tubular shapes. *Context features* can model inter-object relationships and can be used to leverage the information captured by the appearance features adopted.

I adopt a hybrid approach to curvilinear structure segmentation, combining HCFs (used to model the *appearance*) and learned filters (capturing *context* information). In this chapter, I first introduce, derive and assess the detection performance of a new class of (appearance) filters for tortuous and fragmented structures. Then, I discuss and evaluate the detection performance of the methodology to learn *context* filters, aiming to incorporate inter-object relationships and compensate for the modelling limitations of HCFs.

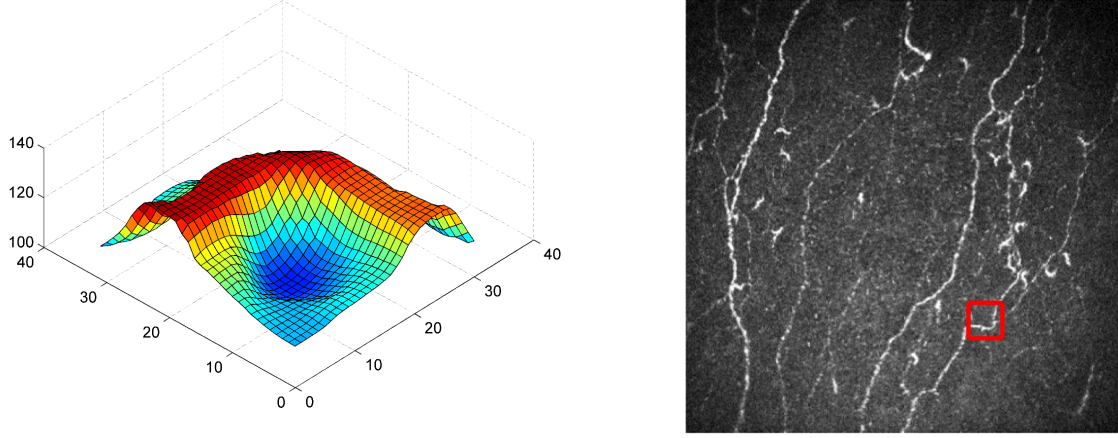


Figure 4.1 The need for a curved-support model. (Left) 3-D profile of the tortuous corneal nerve fibre highlighted with the red rectangle in the original image (right).

4.2 SCIRD: a Scale and Curvature-Invariant Ridge Detector

4.2.1 Curvilinear structure model

The profile of a locally-straight curvilinear structure could be modelled with a Gaussian function. Specifically, let $G(\boldsymbol{\varphi}; \boldsymbol{\sigma})$ be a multivariate, n -D Gaussian function with diagonal covariance matrix, centred at the origin of the coordinate system,

$$G(\boldsymbol{\varphi}; \boldsymbol{\sigma}) = \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^n \sigma_i^2}} \exp\left(-\sum_{i=1}^n \frac{\varphi_i^2}{2\sigma_i^2}\right) \quad (4.1)$$

where $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \dots, \varphi_n)$ represents a point in the $\{\boldsymbol{\varphi}\}$ coordinate system, and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \dots, \sigma_n)$ describes the standard deviation in each direction. A ridge detector can be obtained by measuring the contrast between the part inside and outside the ridge [49]. This can be achieved by combining (e.g. summing, averaging) the responses of the convolution with the second derivatives of the Gaussian function in Equation (4.1)[49].

As shown in Figure 4.1, curvilinear structures in my target data sets (i.e., IVCN100 and IVCN140) often violate the locally-straight assumption and show a curved-support profile

(Figure 4.1 - left), therefore they cannot be modelled with the straight Gaussian function in Equation (4.1). To model such curved-support profile, I apply a non-linear transformation $\mathcal{T} : \mathbb{R}^n \mapsto \mathbb{R}^n$ with $\mathcal{T}(\mathbf{x}) = \boldsymbol{\varphi} = (\varphi_1, \varphi_2, \dots, \varphi_n)$ of the form

$$\varphi_n = x_n + \sum_{i=0}^{n-1} k_{ni} m_{ni}(x_1, x_2, \dots, x_{n-1}) \quad (4.2)$$

and $\varphi_1 = x_1$, where k_{ni} are weights (interpretation below) and the non-linear functions $m_{ni}(x_1, x_2, \dots, x_{n-1})$ have continuous partial derivatives [91].

In the 2-D case ($n = 2$), plugging Equation (4.2) in Equation (4.1) leads to the curved-support bivariate Gaussian function:

$$G(x_1, x_2; \boldsymbol{\sigma}, \mathbf{k}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(x_1 + k_{10})^2}{2\sigma_1^2} - \frac{(x_2 + k_{20}m_{20} + k_{21}m_{21})^2}{2\sigma_2^2}\right), \quad (4.3)$$

where m_{20} and m_{21} depend on x_1 .

If we now consider *quadratic* non-linear functions $m_{ni} = x_i^2$ for $0 < i < n$ and $m_{n0} = 1$, some of the parameters k_{ni} can be regarded intuitively as curvatures [91]. Specifically, I observe that k_{10} controls the elongation *asymmetry* of the shape (i.e. $k_{10} \neq 0$ makes one tail longer than the other), k_{21} its curvature and k_{20} is simply a translation parameter. However, although towards the end-point of a curvilinear structure the local shape would appear asymmetric longitudinally, I set $k_{10} = 0$ to keep the model simple and reduce the amount of free parameters to tune. This seems reasonable since tortuosity estimation, our reference application, would not benefit significantly from a particularly accurate segmentation of end-point regions. Finally, assuming that this model is centred on the specific curvilinear structure of interest, I set $k_{20} = 0$ (this choice will allow us to adopt a simple “max” operator to obtain the ridge detector, as done in the next section).

Therefore, the model I adopt for tortuous and fragmented curvilinear structures is:

$$\Gamma(x_1, x_2; \boldsymbol{\sigma}, \mathbf{k}) = \underbrace{\frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{x_1^2}{2\sigma_1^2}\right)}_{\text{longitudinal}} \underbrace{\exp\left(-\frac{(x_2 + kx_1^2)^2}{2\sigma_2^2}\right)}_{\text{orthogonal}}, \quad (4.4)$$

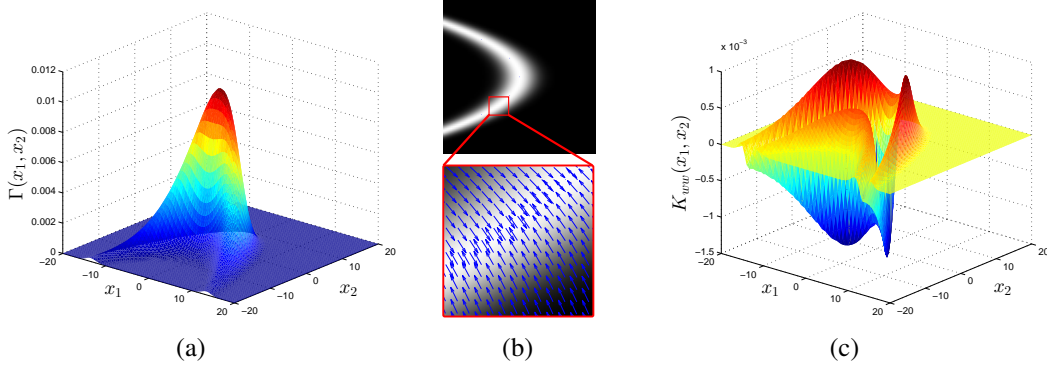


Figure 4.2 *Appearance* model used to capture specific characteristics of tortuous and fragmented structures. (a) An example of our curved-support shape model; (b) the insert (below) visualises the direction along which we measure the second (directional) derivative used to compute the *tubularity probe filter* of the shape model (above), in the window selected; (c) Derived convolutional filter, estimating local tubularity efficiently.

where (x_1, x_2) is a point in the new $\{x\}$ coordinate system of the target structure (orientation assumed known here), (σ_1, σ_2) control the elongation of the shape and its width, respectively; in fact, the first exponential in $\Gamma(x_1, x_2; \sigma, k)$ controls the longitudinal Gaussian profile of the model, and the second the orthogonal one. Unlike previous hand-designed models (e.g., [49, 83, 123]), I add a new parameter, k , to control the curvature of the Gaussian support (Figure 4.2(a)). The benefit of this choice is illustrated by the experimental results in Section 4.2.4.

4.2.2 SCIRD

In this section I derive the ridge detector from the shape model described above.

In order to estimate local tubularity, I adopt a shape-aware measure of contrast between the region inside and outside the curved ridge. Importantly, both the cross-sectional and the longitudinal Gaussian weighting are taken into account, allowing an accurate contrast estimation.

Let $I(x, y)$ be the intensity of a monochromatic image at location (x, y) in image coordinates (the same derivation could be applied to each channel of a colour image). To

make the contrast measure shape-aware locally, I introduce local first-order gauge coordinates $\{\mathbf{v}(x, y), \mathbf{w}(x, y)\}$, where $\mathbf{w}(x, y) = \frac{\nabla I(x, y)}{\|\nabla I(x, y)\|}$ and $\mathbf{v}(x, y) = \mathbf{w}_\perp(x, y)$ ¹.

A shape-aware contrast measure is then given by the second directional derivative in the direction orthogonal to the model centreline:

$$I_{\mathbf{w}\mathbf{w}} = D_{\mathbf{w}}[D_{\mathbf{w}}I] = D_{\mathbf{w}}[\mathbf{w}^\top \nabla I] \triangleq \mathbf{w}^\top H_I \mathbf{w}, \quad (4.5)$$

where $D_{\mathbf{w}}$ is the directional derivative operator along \mathbf{w} ,

$$H_I = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{bmatrix} \quad (4.6)$$

is the Hessian matrix, and $\nabla I = [I_x, I_y]^\top$ is the gradient of I in x - y coordinates. Substituting Equation (4.6) in Equation (4.5) we obtain:

$$I_{\mathbf{w}\mathbf{w}} = \frac{(I_x I_{xx} + I_y I_{yx})I_x + (I_x I_{xy} + I_y I_{yy})I_y}{I_x^2 + I_y^2}, \quad (4.7)$$

where I have omitted arguments $(x, y; \boldsymbol{\sigma}, k)$ for compactness. We can differentiate by convolving the image with derivatives of the curved-support Gaussian, which leads to an efficient tubularity estimation based on the convolution with a filter bank that *can be pre-computed off-line*:

$$I_{\mathbf{w}\mathbf{w}}(x, y; \boldsymbol{\sigma}, k) = I(x, y) * K_{\mathbf{w}\mathbf{w}}(x, y; \boldsymbol{\sigma}, k), \quad (4.8)$$

where $K_{\mathbf{w}\mathbf{w}}$ represents our *tubularity probe kernel* (see example in Figure 4.2(c)):

$$K_{\mathbf{w}\mathbf{w}} = \frac{(\tilde{\Gamma}_x \Gamma_{xx} + \tilde{\Gamma}_y \Gamma_{yx})\tilde{\Gamma}_x + (\tilde{\Gamma}_x \Gamma_{xy} + \tilde{\Gamma}_y \Gamma_{yy})\tilde{\Gamma}_y}{\tilde{\Gamma}_x^2 + \tilde{\Gamma}_y^2}. \quad (4.9)$$

Here, $\tilde{\Gamma}(x_1, x_2; \boldsymbol{\sigma}, k)$ is a curved-support Gaussian model with a constant (hence non-Gaussian) longitudinal profile; its gradient direction is orthogonal to the centreline (Fig-

¹The symbol \perp denotes orthogonality.

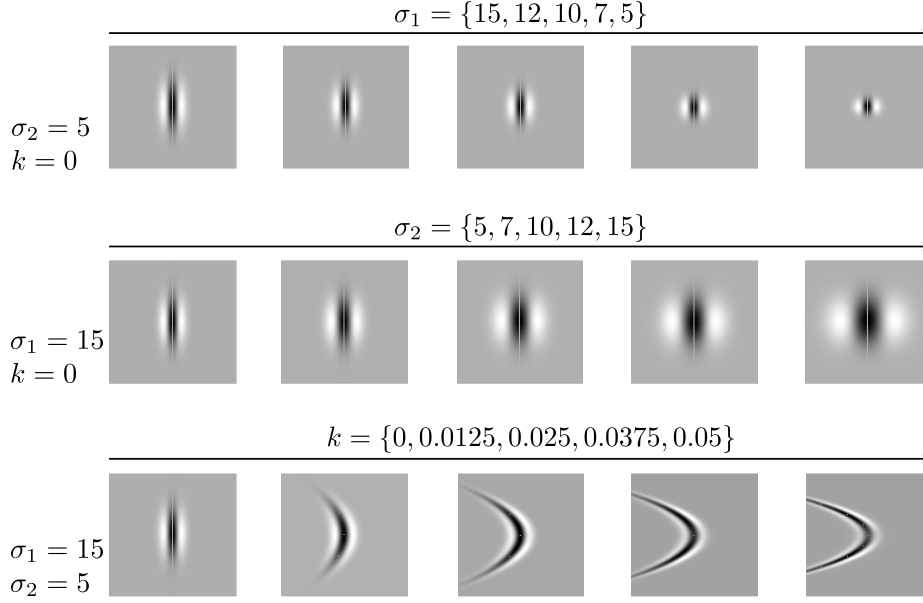


Figure 4.3 The effect of geometric parameters on the shape of SCIRD filters.

ure 4.2(b)):

$$\tilde{\Gamma}(x_1, x_2; \boldsymbol{\sigma}, k) \propto \exp\left(-\frac{(x_2 + kx_1^2)^2}{2\sigma_2^2}\right). \quad (4.10)$$

To achieve invariance in the discrete domain, I create a filter bank of $K_{\text{ww}}(x, y; \sigma_1, \sigma_2, k, \theta)$ kernels generated (notice, *off-line* for efficiency) by making σ_2 (i.e., scale) and k (i.e., curvature) span the respective ranges, $\{\sigma_2^{(i)}, i = 1, \dots, N_{\sigma_2}\}$ and $\{k^{(i)}, i = 1, \dots, N_k\}$, in the curvilinear structures of interest. These ranges are easily established by image inspection and remain valid, in my experiments, for all images of a given type (e.g., corneal fibres, dendrites). To estimate local orientation (assumed known in Equation 4.4) I expand the filter bank with kernel replicas rotated by $\theta \in [0, 2\pi)$, $\{\theta^{(i)}, i = 1, \dots, N_\theta\}$. Finally, I account for fragmented (point-like) structures by adopting a suitable range for σ_1 , $\{\sigma_1^{(i)}, i = 1, \dots, N_{\sigma_1}\}$. As above, this range is easily established by direct inspection. Figure 4.3 shows the effect of the main parameters to control the shape of the proposed convolutional tubularity kernels.

Once the filter bank has been precomputed, approximate² scale, curvature, rotation, and elongation invariance is obtained by maximising the tubularity measure $I_{\mathbf{ww}}(x, y; \sigma_1, \sigma_2, k, \theta)$ for each pixel $(x, y) \in \mathcal{D}(I)$ across the pre-defined parameter space $\{(\sigma_1, \sigma_2, k, \theta)^{(j)}, j = 1, \dots, N_{\sigma_1} + N_{\sigma_2} + N_k + N_\theta\}$:

$$\left(\sigma_1, \sigma_2, k, \theta\right)^* = \arg \max_{\sigma_1, \sigma_2, k, \theta} I_{\mathbf{ww}}(x, y; \sigma_1, \sigma_2, k, \theta). \quad (4.11)$$

Notice that, by definition, the maximum value of $I_{\mathbf{ww}}(x, y; \sigma_1, \sigma_2, k, \theta)$ at each pixel across the parameter space corresponds to the measure of tubularity we defined in Equation (4.8). Since our tubularity measure is based on contrast, its sign, computed on tubular objects, depends on whether the inside or the outside region is brighter. When $I_{\mathbf{ww}}$ is expected to be positive, for instance, negative local tubularity responses are likely due to a non-target objects and can be discarded. To this aim, I apply thresholding such that $I_{\mathbf{ww}}^T = \max(I_{\mathbf{ww}}, 0)$.

In some cases images show significant intra- and inter-image contrast variability. This is the case, for instance, for corneal nerve fibres in confocal microscopy images (Figure 3.1), due to illumination and the reflectance of the ocular tissue. In these images, local tubularity estimation tends to be biased towards high contrast, penalising thin and poorly contrasted corneal nerve fibres. To alleviate this problem and achieve the desired level of contrast invariance, I introduce a contrast normalization term with a parameter $\alpha \in \mathbb{R}$ in the proposed ridge detector, emphasising the response of low-contrast structures:

$$SCIRD(x, y; \{\sigma_1, \sigma_2, k, \theta\}^*) = \frac{I_{\mathbf{ww}}^T(x, y; \{\sigma_1, \sigma_2, k, \theta\}^*)}{1 + \alpha I^C(x, y)}, \quad (4.12)$$

where

$$I^C(x, y) = \frac{1}{(N+1)^2} \sum_{i=x-\frac{N}{2}}^{x+\frac{N}{2}} \sum_{j=y-\frac{N}{2}}^{y+\frac{N}{2}} \max_{\sigma_2} \|\nabla I(i, j; \sigma_2)\|_2 \quad (4.13)$$

²A theoretical invariance could be obtained by introducing normalised coordinates and deriving the normalising terms for each filter, as shown by Lindeberg [92] in the simple case of symmetric Gaussian functions.

is the adopted contrast measure based on multiscale gradient magnitude estimation averaged on a $(N + 1) \times (N + 1)$ patch around pixel $(x, y) \in \mathcal{D}(I)$. Notice, N is *not* a free parameter, but the width (and height) of the largest filter in the SCIRD filter bank ($N = 8\sigma_{2max}$).

Based on the experimental validation in Section 4.2.4, SCIRD achieves good detection performance (outperforming state-of-the-art HCFs). However, segmentation and centreline detection can be improved by introducing a supervised classifier, as discussed in the next section.

4.2.3 Supervised SCIRD

The selection of a single filter response (i.e. the maximum) over all the filters does not exploit the discriminative power of the entire filter bank. On the other hand, finding the optimal (in general, non-linear) combination and weighting factor of each filter is not trivial. Here, I tackle this problem by proposing the supervised version of SCIRD. To this aim, I combine $n_S = N_{\sigma_1} + N_{\sigma_2} + N_k + N_\theta$ feature maps $(I_{ww}^{(i)}, i = 1, \dots, n_S)$ obtained using our filter bank with SCIRD to form a feature vector \mathbf{f} :

$$\mathbf{f} = \left[\text{SCIRD}, I_{ww}^{(1)}, I_{ww}^{(2)}, \dots, I_{ww}^{(n_S)} \right]^\top, \quad (4.14)$$

and use it to classify each pixel. I employ a Random Forest [22, 39] as a classifier, due to its better performance (at parity of feature vector and training protocol) as compared with other classifiers, e.g. support vector machines or SVM [37], on several tasks [39], including curvilinear structure segmentation [111]. Thresholding directly the resulting probability map (which can be seen as a tubularity map) leads indeed to a more accurate and less noisy segmentation, as confirmed by the experiments in Section 4.2.4. Supervised centreline detection is obtained using pixel-wise non-maxima suppression and thresholding on the tubularity map. As local orientation for both supervised and unsupervised centreline detection I choose that of the kernel responding maximally.

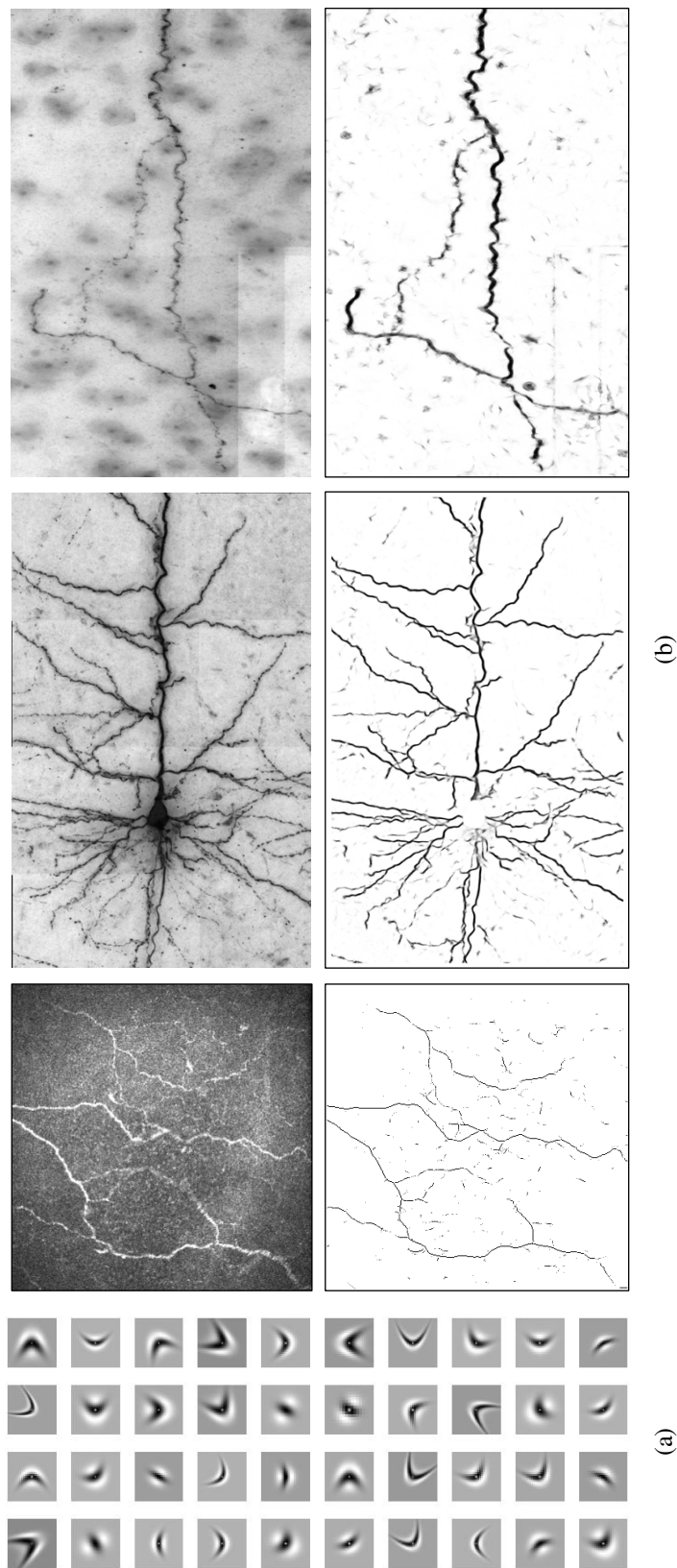


Figure 4.4 (a) A subset of SCIRD filters used in our experiments. Notice, the model includes straight support too; (b) original images (top), supervised SCIRD *tubularity* maps (bottom) from IVCN (left), BF2D (centre), VC6 (right) testing sets.

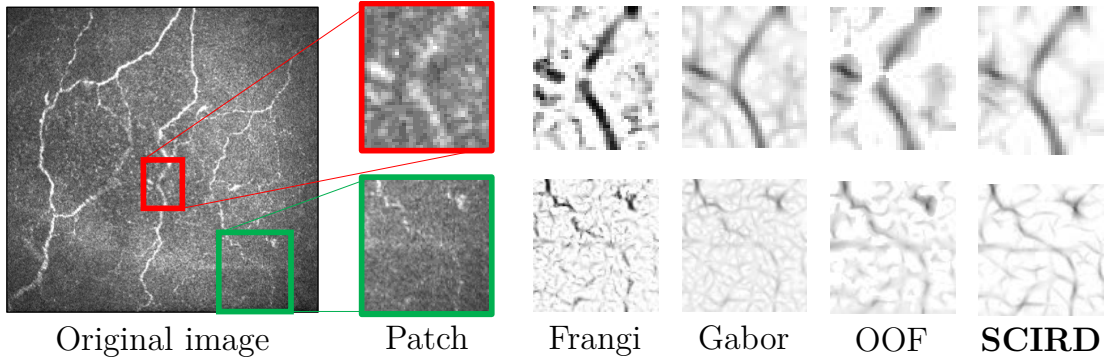


Figure 4.5 Qualitative comparison: tubularity estimation results on IVCM. SCIRD (our approach) shows better connectivity and higher signal-to-noise ratio than others.

4.2.4 Experiments and results

Data sets. I validate SCIRD segmentation performance on IVC100, BF2D and VC6 described in Section 3.2. These datasets include low and high resolution images of corneal nerve fibres and neurons, showing a wide range of tortuosity characteristics as illustrated in Figure 4.4(b). Notice that the task to be performed on IVC100 is actually centreline detection (our collaborators traced only corneal nerve fibre centrelines). As usually done to evaluate methods extracting one-pixel-wide curves (e.g. by Sironi et al. in [119, 122]), I introduce a tolerance factor ρ : a predicted centreline point is considered a true positive if it is at most ρ distant from a ground truth centreline point. Following Sironi et al. [119, 122], $\rho = 2$ pixels in these experiments. BF2D and VC6 include a training and test set, while IVC100 does not. Therefore, for IVC100, I average performance measures over 10 random sub-sampling cross-validation runs, using 50 images for training and the rest for testing in each run, following the established benchmarking procedure (e.g. [111]). The resulting precision-recall curves are reported in Figure 4.6 (mean and standard deviation of the results from individual runs).

Parameters setting. SCIRD’s key parameters are σ_1 , σ_2 and k controlling the filter elongation, width and curvature, respectively. These parameters have been set by visual inspection, following previous performance evaluation protocols for HCFs [49, 83, 111, 123]. The driving idea is to set the parameters such that the shape of SCIRD filters resembles the one of actual curvilinear structures observed in the data set. Specifically, the ranges of σ_1

over data sets were chosen considering the level of fragmentation (i.e. the more fragmented, the wider the ranges). I set σ_2 values based on the maximum and minimum width of the target structures in each dataset as they depend on resolution; curvature values were set according to the level of tortuosity (or bending) shown by the specific curvilinear structures. The discretisation step of the orientations θ was set to achieve a good trade-off between accuracy and speed. The contrast normalisation parameter α was set to achieve the desired level of contrast invariance.

I set a single range (sampled with fixed discretisation step³) for θ and k for all the datasets: $\theta = \{\frac{\pi}{12}, \frac{\pi}{6}, \dots, 2\pi\}$, and $k = \{0, 0.025, \dots, 0.1\}$. I set other parameters separately, given the significant difference in resolution between data sets. In particular, for the IVC100, $\sigma_1 = \{2, 3, 4\}$, $\sigma_2 = \{2, 3\}$, $\alpha = 1$; for the BF2D dataset, $\sigma_1 = 5$, $\sigma_2 = \{2, 3, 4\}$, $\alpha = -0.075$; for the VC6, $\sigma_1 = 3$, $\sigma_2 = \{2, 3\}$, $\alpha = 0$. For the supervised SCIRD, I used the same range for θ and k , but I doubled the discretisation steps for computational efficiency.

To provide a fair comparison, parameters for SCIRD and baseline methods were manually tuned on training data to achieve their best performance on each dataset.

I trained Random Forests using 50,000 pixels randomly selected from the training set of each data set used in these experiments. The number of decision trees and maximum number of samples in each leaf was set using the out-of-bag error [22].

Results and discussion. I compare SCIRD with three hand-crafted ridge detectors: Frangi [49], Gabor [123], and the recent Optimally Oriented Flux (OOF) [83]. These methods form a representative set of state-of-the-art and accurate detectors of tubular structures. Qualitative (Figure 4.5) and quantitative (Figure 4.6) results show that SCIRD outperforms the baselines considered on all datasets. Specifically, SCIRD shows higher precision from medium to high recall values for the IVC100 dataset, suggesting that our filters behave better than others at low resolution and low SNR when dealing with tortuous and fragmented structures. The low number of false positives when the number of false negatives is low implies that SCIRD selects target structures with higher confidence. Notice that contrast en-

³Notice that using regularly spaced samples on a grid as done here (and by other authors in the context of curvilinear structure segmentation [49, 83, 111, 123]) is not an optimal sampling strategy. Later, in Section 6.3.1, I will use a better sampling approach based on k-means (locally optimal).

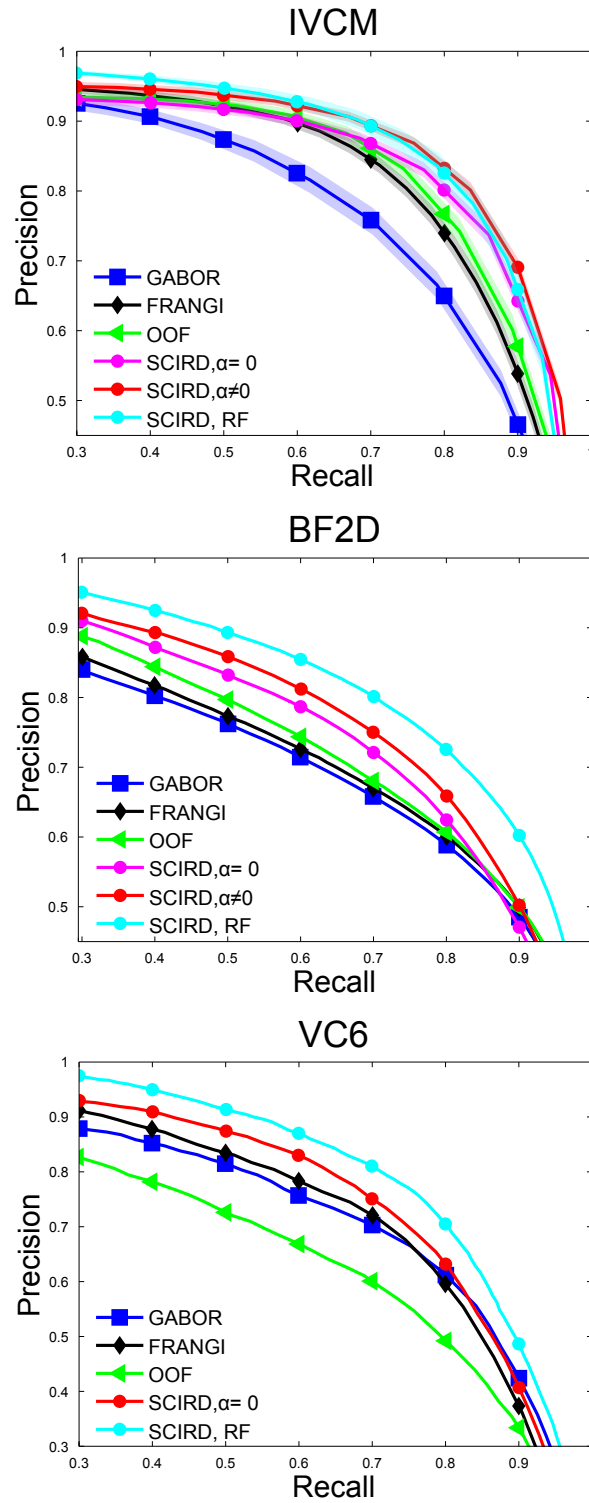


Figure 4.6 Precision-recall curves for SCIRD and baselines on IVCM, BF2D and VC6 datasets. Curves are obtained applying different thresholds on the *tubularity* maps.

hancement (“SCIRD $\alpha \neq 0$ ”, $\alpha = 1$) boosts performance on this dataset achieving the level of performance obtained in the supervised setting (“SCIRD, RF”), at a lower computational cost. For the BF2D dataset, SCIRD shows a significant improvement from low to high recall values, suggesting that fewer non-target structures are detected and targets are enhanced with higher accuracy (e.g. point-like structures are correctly reconnected, tortuous structure profiles are better preserved). Contrast reduction (“SCIRD $\alpha \neq 0$ ”, $\alpha = -0.075$) proves helpful for these images. Supervised classification improves performance further. For the VC6 dataset, SCIRD shows better performance from low to medium recall values, indicating a better discrimination between curvilinear structures and artifacts, in addition to a more accurate profile segmentation for tortuous structures. Contrast enhancement did not help for this dataset, while supervised classification contributes substantially to improve results.

The time to run SCIRD using a single core on an Intel i7-4770 CPU @ 3.4 GHz and MATLAB code (R2014a) is 3.84s (IVCM), 2.77s (VC6) and 23s (BF2D). The structure of the SCIRD algorithm is highly parallelizable (filter bank pre-computed off-line), which could lead to dramatic speed-ups on a parallel architecture.

Although SCIRD, especially the supervised version, shows good detection performance on the three data sets used for validation, it does not exploit context information (i.e. inter-object relationships) found to improve segmentation performance in other applications [1, 99, 133]. Modelling context using HCFs is challenging due to the variety of configurations which should be considered. In the next section, I will discuss how to complement SCIRD (HCFs, in general) with learned context filters to leverage context information efficiently.

4.3 Learning context filters

As discussed in Section 2.2.3, feature extraction based on hybrid methods is particularly appealing as exploiting the efficiency of fast HCFs while limiting the amount of learned filters. Rigamonti et al. [111] proposed to combine learned *appearance* filters with HCFs *also* modelling *appearance* (i.e., OOF). However, modelling and including *context* information in a segmentation framework has been recently shown to outperform considerably solutions

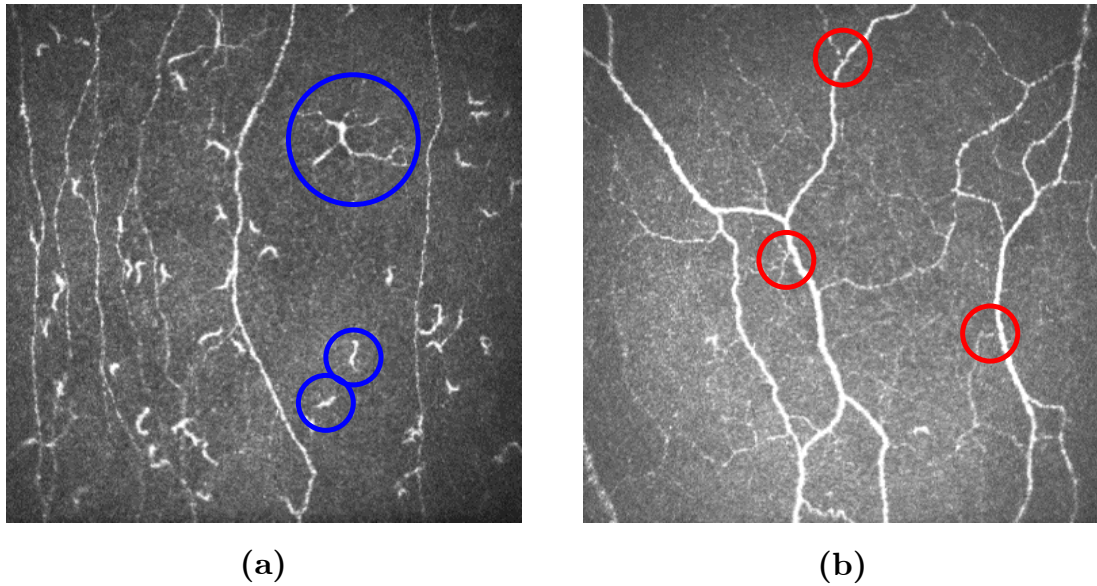


Figure 4.7 Visual examples motivating the need for including *context* information in the segmentation pipeline. (a - blue circles) short, well contrasted, but isolated segments, which could be wrongly segmented as corneal nerve fibres (they are dendritic cells, instead); and (b - red circles) short and poorly contrasted segments branching from corneal nerve fibres, which should be segmented as corneal nerve fibres.

based only on *appearance* features (see, for instance, the auto-context framework proposed by Tu and Bai [133] and the spin-context algorithm by McKenna et al. [99]). I build on the same idea here.

Many types of information can be referred to as *context*[133]: different parts of an object can be context to each other; different objects in an image can be each other's context. For example, a short, well contrasted, but isolated segment may suggest it is not a corneal nerve fibre (but rather a dendritic cell due to inflammation, as shown in Figure 4.7(a)). Instead, a short and poorly contrasted segment branching from a clearly visible corneal nerve fibre may suggest that the segment is a fibre (Figure 4.7(b)). Learned (multi-range) context filters presented here, represent an alternative and efficient solution to include context information in a segmentation framework, which is experimentally shown to help in ambiguous situations such as the ones shown in Figure 4.7.

In Section 4.3.1 I summarise the algorithm adopted for unsupervised filter learning. Section 4.3.2 and 4.3.3 describe how to use this algorithm to learn context filters efficiently and

then combine *appearance* and *context* information using a *single* (hence fast) discriminative model. Experiments and results for single-range context filters are reported and discussed in Section 4.3.4.

4.3.1 Unsupervised filter learning

Although K -means is not designed to learn sparse representations, such as SC or independent component analysis, experimental results [36] suggest that it tends to learn sparse projections of the data under two assumptions: (1) a sufficiently large amount of training data, given the data dimensionality; (2) applying whitening as pre-processing to remove correlations between data components. Both hold in my experimental setting, so that I can employ K -means clustering to learn context filters instead of more expensive algorithms such as SC [111]. Here, I adopt the algorithm by Coates et al. [36] which I summarise concisely.

The goal is to learn a dictionary $D \in \mathbb{R}^{q \times K}$ of K vectors so that a data vector $\mathbf{x}^{(i)} \in \mathbb{R}^q, i = 1, \dots, m_D$ can be mapped to the code vector that minimizes the reconstruction error. Before running the learning algorithm I normalize the brightness and contrast of each input data point $\mathbf{x}^{(i)}$, in this case $p \times p$ patches. Then, I apply patch-level whitening through the ZCA transform [36] so that $\mathbf{x}_{ZCA}^{(i)} = V(\Sigma + \epsilon_{ZCA}\mathbb{I})^{-1/2}V^\top \mathbf{x}^{(i)}$, where V and Σ are computed from the eigenvalue decomposition of the data points covariance $V\Sigma V^\top = \text{cov}(X)$, and ϵ_{ZCA} is a small constant controlling the trade-off between whitening and noise amplification. After pre-processing the patches, we solve the optimization problem

$$\underset{D, \mathbf{c}}{\text{argmin}} \sum_i \left\| D\mathbf{c}^{(i)} - \mathbf{x}_{ZCA}^{(i)} \right\|_2^2 \quad (4.15)$$

subject to $\|\mathbf{c}^{(i)}\|_0 \leq 1, \forall i = 1, \dots, m_D$ and $\|\mathbf{d}^{(j)}\|_2 = 1, \forall j = 1, \dots, K$, where $\mathbf{c}^{(i)}$ is the code vector related to input $\mathbf{x}_{ZCA}^{(i)}$, and $\mathbf{d}^{(j)}$ is the j -th column of the dictionary D .

Combining appearance filters learned through K -means (or SC) with HCFs leads to the combination approach proposed in [111] based on appearance-only features (see Figure 4.8 left).

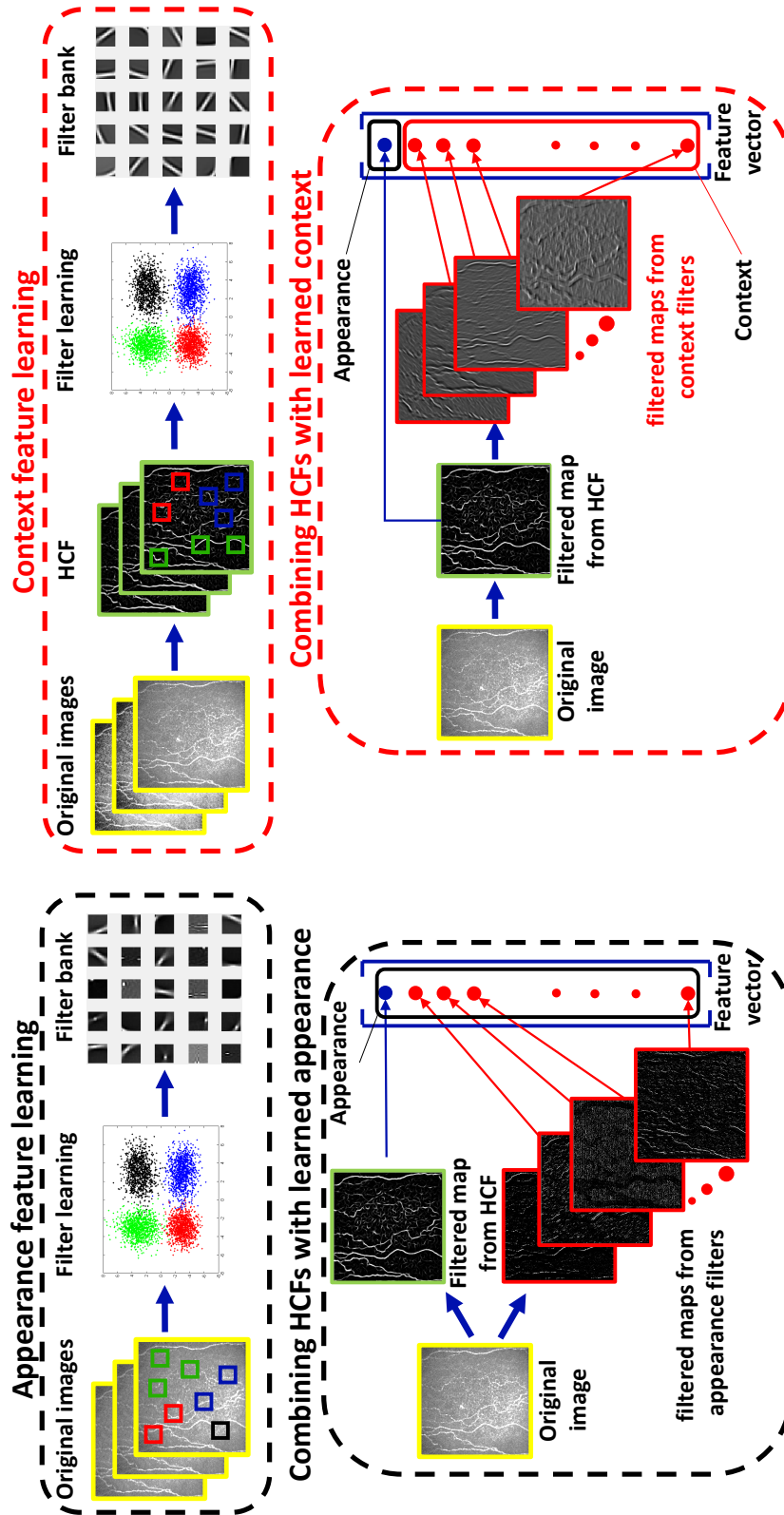


Figure 4.8 Difference in learning appearance and context filters. (Left) Appearance filters are learned using original image patches and applied on the same layer of the HCFs, thus leading to potential redundancy in the feature set. (Right) Context filters are learned using HCF maps and applied after the HCFs, thus eliminating redundancy.

4.3.2 Context filters

Hand-crafted and learned appearance features are designed to capture object-specific properties. In the case of curvilinear structures, they detect characteristics that make them *appear* as ridge-, vessel- or tube-like shapes. However, appearance features do not take into account specific inter-object relationships, that is, *context* information that has been recently shown to improve performance substantially in segmentation tasks over methods employing appearance features only. A well-known method including context information is auto-context [133], which learns multiple discriminative models sequentially. This imposes an extra computational cost over learning a single discriminative model, as in traditional methods based on features representing object pixels and a single classifier to infer their labels. While an extra computational cost may have little impact on training/testing time for applications involving small datasets or small images, it may become impractical with large volumes of image data. This is the case, for instance, with large healthcare screening programs based on images, e.g., diabetic retinopathy [67]. For this reason, I aim to include context information in a learning method without increasing the computational cost with respect to the solution proposed by Rigamonti and Lepetit [111]. This is achieved by learning a *single* discriminative model which takes as input both *appearance* (i.e. likelihood computed on the original image) and, unlike Rigamonti and Lepetit’s method [111], *context information* (i.e. relations between objects). To model appearance, here I employ the fast Optimally Oriented Flux (OOF) feature [83], shown to outperform other HCFs on the datasets I use for validation[111]⁴. I include context information by learning context filters to be used in combination with the HCFs in a new hybrid model to segment curvilinear structures.

Learning context filters has two clear advantages: 1) including higher-level information in a hybrid framework; 2) high efficiency and adaptability since convolution with filter banks is very fast even on standard computers. In addition, the proposed method has a key advantage over methods learning appearance filters as proposed by Rigamonti and Lepetit [111]: it implicitly eliminates, or reduces noticeably, the redundancy of learned filters. In fact, learned appearance filters may be reconstructed through a combination (linear or

⁴Experiments using SCIRD as base HCF are reported later in the chapter.

non-linear) of the HCFs already used to model appearance. Figure 4.8 shows the difference between the proposed approach (right) and the combination method proposed in[111] (left). Notice, while appearance filters are learned on the same layer where HCFs are applied (original image), context filters are learned on a *different* layer (i.e. after HCFs are applied).

4.3.3 Description vector and supervised classification

I apply learned filters to the input image to compute multiple feature maps efficiently using correlation:

$$\mathbf{L}^{(j)} = \mathbf{D}^{(j)} \circ \mathbf{I}_n, \quad (4.16)$$

where $\mathbf{D}^{(j)}$ is the j -th learned filter, \mathbf{I}_n is the normalized input image (i.e., zero mean and unit standard deviation) and the symbol \circ denotes correlation. I have also experimented with normalizations of the input image at patch level (including whitening), measuring the squared distance between the pre-processed patch and each filter; unreported experiments show that these normalizations, although important during the filter learning procedure, do not improve performance noticeably and increase the computational cost. Thus, for each image location (u, v) , I construct the following description vector:

$$\left[\underbrace{\text{OOF}(u, v)}_{\text{appearance}}, \underbrace{\mathbf{L}^{(1)}(u, v), \dots, \mathbf{L}^{(N)}(u, v)}_{\text{context}} \right]^T, \quad (4.17)$$

including appearance and learned context features ($N \leq 200$ in my experiments). I then apply a Random Forest to classify each pixel, for the same reasons detailed above. Centreline detection is obtained using pixel-wise non-maxima suppression and thresholding on the tubularity map. Local orientation is estimated using OOF.

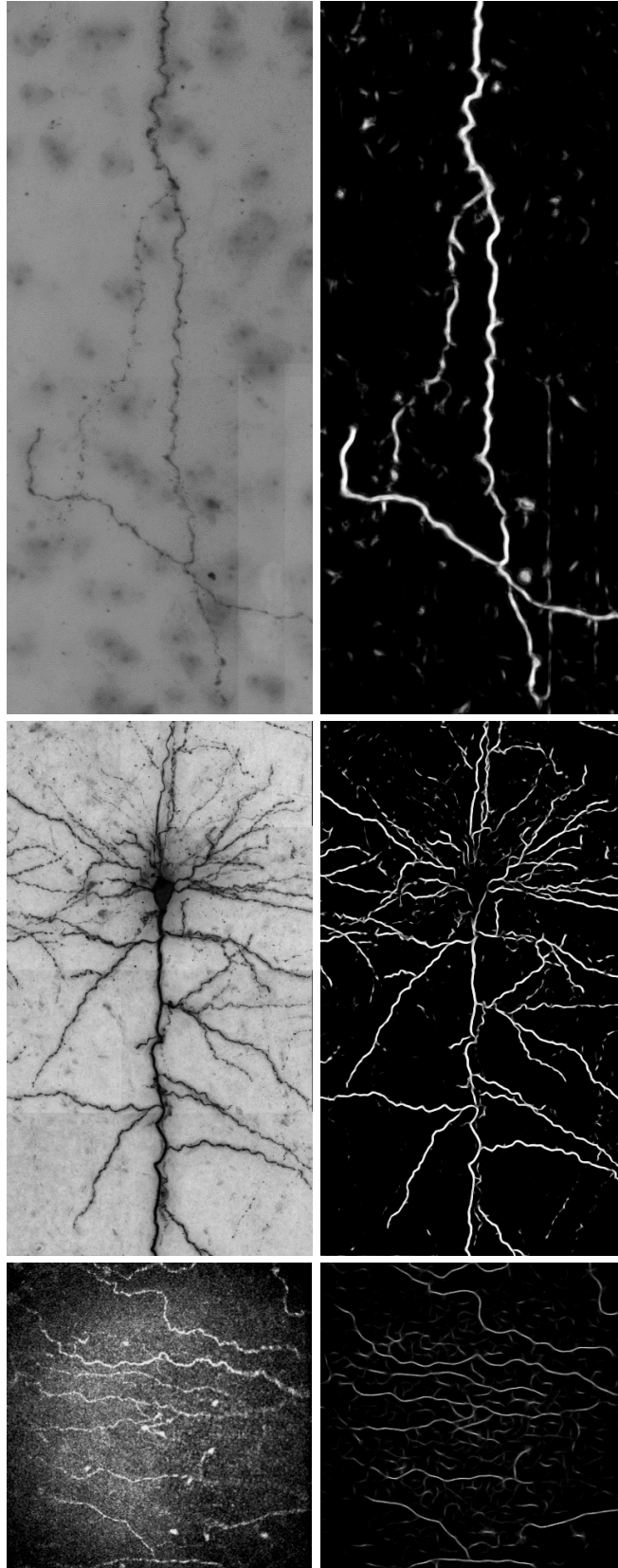


Figure 4.9 Original images (top) and tubularity maps (bottom) obtained with our approach on IVC6 (left), BF2D (centre), VC6 (right) datasets.

4.3.4 Experiments and results

Datasets. I validate the proposed combination method using the same three data sets and protocol used to validate SCIRD. Some qualitative results are shown for illustration in Figure 4.9.

Parameter setting. First, images are normalised to have zero mean and unit standard deviation. When HCFs are used as baselines, all parameters are tuned for each dataset independently to achieve best performance on each data set [111]. To reduce the number of parameters to be optimised over datasets and test generalization, I fixed the OOF parameter ranges, the whitening parameter ϵ_{ZCA} and the filters' size in the filter banks to the same values for *all* datasets. For OOF, I set $\sigma = \{2, 3, 4\}$ (Eq. (8) in [83]) and $R = \{2, 3, 4\}$ (Eq. (5) in [84]). ϵ_{ZCA} was set to 0.001, considering the trade-off between noise amplification and filters sharpness[36]. Filters' size was set to 11×11 pixels. Notice that the chosen patch size allows us to collect a sufficient number of patches to learn dictionaries, in agreement with the guidelines in [36]⁵. I adopted filter banks of 100 filters (i.e., $N = 100$) as a good compromise between accuracy and speed. I used Random Forests with 100 random trees to achieve fast predictions. I trained classifiers using the same number of positive and negative samples and priors estimated empirically. Experiments were run on an 8-core 64-bit architecture using MATLAB (2014a) implementations.

Results and Discussion. I compare the proposed method, combining appearance and context filters, with the one recently introduced by Rigamonti and Lepetit in [111] based only on appearance. Therefore, I report results obtained with the original implementation of [111] using 9 appearance filters learned through SC ("*Rigamonti et al. [8]*" in Figure 4.6)⁶. I also report experiments using k-means to learn appearance filters ("*OOF, learned appearance*") and the same dictionary size I adopted to learn context filters.

As Figure 4.10 shows, the proposed combination method ("*OOF, learned context*") outperforms the baselines on all datasets, especially on VC6 and IVCN (on a rather large range of recall values). The different datasets allow us to compare performance with very di-

⁵I selected 340,000 patches in the worst case represented by a single training image of the BF2D dataset. Notice that, in [36] 100,000 16×16 patches are considered sufficient.

⁶Original code at: https://bitbucket.org/roberto_rigamonti/med_img_pc.

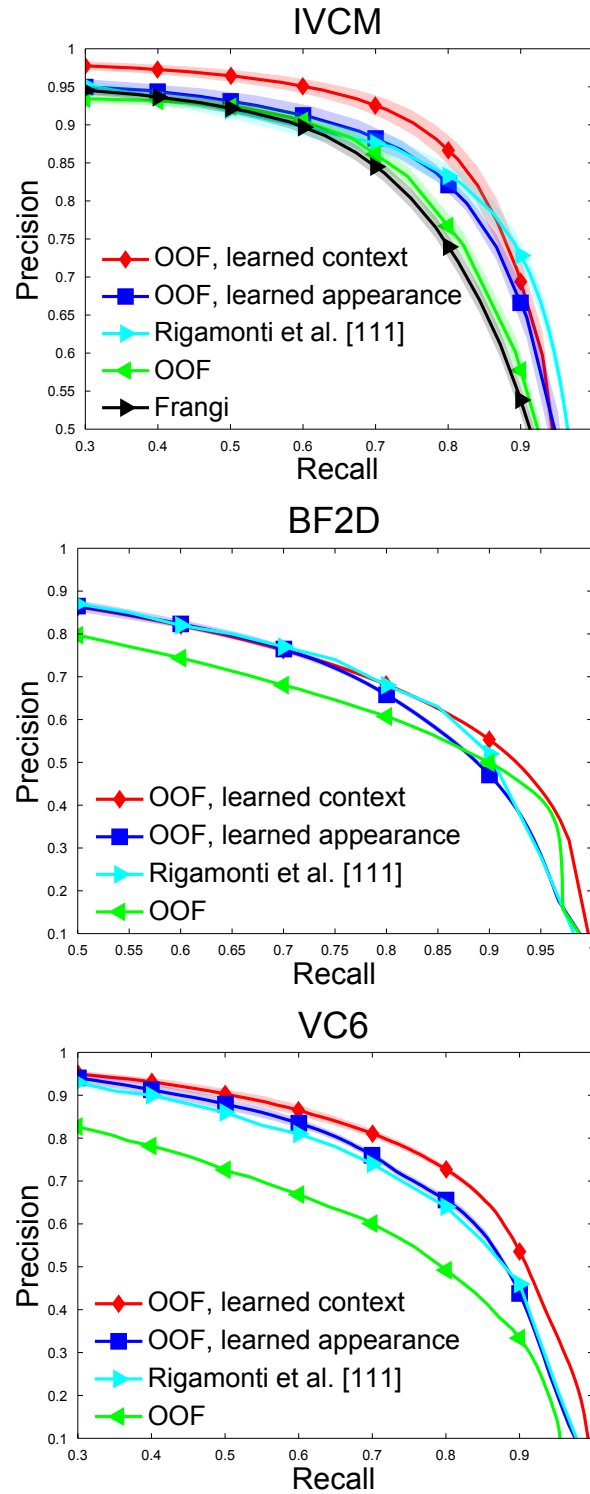


Figure 4.10 Precision-recall curves for pixel-level classification. Shaded color bands represent 1 standard deviation of the results from individual runs.

verse data characteristics: with tortuous, fragmented structures and low signal-to-noise ratio (IVCM), the proposed method shows a better precision for low recall values as it better segments fragmented structures; when structures have better contrast but point-like appearance (BF2D), the proposed method shows higher precision at high recall values as it reduces the false positives due to point-like non-target structures and increases the connectivity; when dealing with complex non-target structures (e.g. blobs in VC6), the proposed approach shows better performance from medium to high recall values, since it reduces the amount of false positives due to such structures.

Since my main goal is to compare learned context with learned appearance regardless of the learning algorithm used (k-means or SC), I assess the effect of patch and dictionary size on the performance measured using AUPRC for both combination methods using k-means as learning method. As Table 4.1 shows, combining OOF with learned context (proposed here) outperforms the combination with learned appearance [111] in terms of AUPRC regardless of the chosen patch and dictionary size. Moreover, learning as little as 10 context filters gives the same or even better performance than learning 100 appearance filters, thus confirming that the proposed approach reduces potential redundancy.

As expected, learning appearance filters using convolutional SC [111] improves performance over k-means (at a parity of dictionary size), at a price of a disproportionate decrease in speed. Modest improvements (AUPRC are 0.8748 vs 0.8557 on IVCM, 0.7700 vs 0.7460 on VC6, 0.7955 vs 0.7857 on BF2D) are achieved with a speed loss of two orders of magnitude (speed 25-30 mins vs a few seconds) using the same machine. However, learning 10 context filters with k-means yields better AUPRC than Rigamonti and Lepetit's method [111] (AUPRC are 0.8866 vs 0.8748 on IVCM, 0.7911 vs 0.7700 on VC6, 0.7966 vs 0.7955 on BF2D). Also, while learning with SC is time-consuming for filter banks larger than 100 (several days are reportedly needed to learn a filter bank of 121 filters [111] on machines comparable to the ones used in my experiments), a few minutes are required to learn as many as 200 context filters with k-means, in case a larger dictionary is needed (e.g. for VC6). As a result, the proposed method combining OOF with context filters learned using

Table 4.1 Effect of patch and dictionary size on the area under precision-recall curves (mean/standard deviation). Our method outperforms the baseline in all conditions.

IVCM	Patch size (pixels)			Number of learned filters (N)		
	11×11	15×15	21×21	10	100	200
OOF, learned context	0.8928/0.0056	0.9078/0.0043	0.9133/0.0052	0.8866/0.0033	0.8928/0.0056	0.8976/0.0037
OOF, learned appearance	0.8665/0.0114	0.8878/0.0059	0.8870/0.0102	0.8557/0.0069	0.8665/0.0114	0.8878/0.0039
	100 learned filters			Patch size: 11×11 pixels		
BF2D	Patch size (pixels)			Number of learned filters (N)		
	11×11	15×15	21×21	10	100	200
OOF, learned context	0.8057/0.0017	0.8010/0.0029	0.7948/0.0021	0.7966/0.0043	0.8057/0.0017	0.8054/0.0022
OOF, learned appearance	0.7881/0.0060	0.7888/0.0035	0.7908/0.0036	0.7857/0.0026	0.7881/0.0060	0.7824/0.0044
	100 learned filters			Patch size: 11×11 pixels		
VC6	Patch size (pixels)			Number of learned filters (N)		
	11×11	15×15	21×21	10	100	200
OOF, learned context	0.8272/0.0052	0.8295/0.0035	0.8069/0.0063	0.7911/0.0070	0.8272/0.0052	0.8368/0.0041
OOF, learned appearance	0.7905/0.0063	0.7904/0.0045	0.7809/0.0032	0.7460/0.0053	0.7905/0.0063	0.7905/0.0062
	100 learned filters			Patch size: 11×11 pixels		

k-means outperforms substantially the method proposed in [111]: the best AUPRC figures are 0.9078 vs 0.8748 on IVCM, 0.8368 vs 0.7700 on VC6, 0.8057 vs 0.7955 on BF2D.

4.4 Learning *multi-range* context filters

The approach above does capture some level of context, but it has limitations. First, the *range* p (in pixels) of spatial context captured in each direction is dictated by the patch size ($p \times p$). This parameter can be set by cross-validation, but its maximum value p_M is limited by the amount of available training data (see assumption (1) above on applying K -means). Notice that the data points dimensionality $q \sim p^2$, thus forcing p_M to be relatively small (e.g., less than 30 pixels) if training data is not abundant (e.g., less than 500,000 patches), a typical case in several medical applications. So, this approach fails to model *long-range* context information. Second, it is a *single-range* context model, i.e. it captures inter-object relationships characterising a specific neighbourhood only.

Here, I address these issues by introducing a *multi-scale* architecture of learned context filters modelling *multi-range* spatial context (Section 4.4.1), the benefits of which are assessed and discussed in Section 4.4.2.

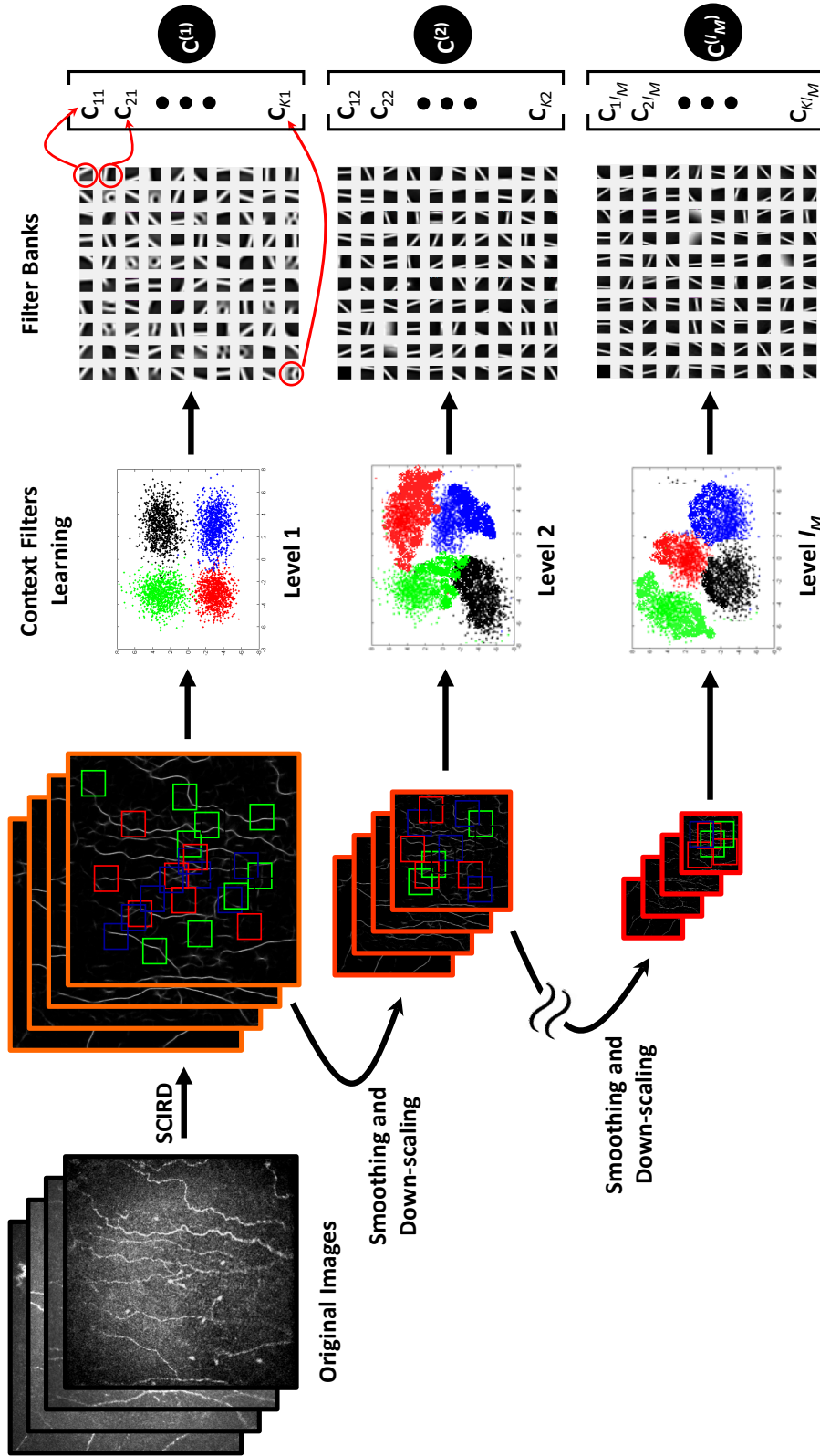


Figure 4.11 Block diagram of the proposed unsupervised *multi-range* context filters learning. Notice, we keep the patch size constant over the levels to capture larger and larger spatial context (light orange, orange and red indicate short-, medium- and long-range context, respectively).

4.4.1 *Multi-range* context filters

Let us denote with $\{I_T^{(S_1)}\}$ the set of tubularity maps obtained after SCIRD (or any other HCF) is applied to the training images $\{I_T\}$, where $T = 1, \dots, T_M$. The proposed *multi-range* context filters architecture is obtained by learning filters on $\{I_T^{(S_1)}\}$ and their smoothed, downsampled versions $\{I_T^{(S_{\mathcal{L}})}\}$, where $\mathcal{L} = 1, \dots, l_M$ indicates a certain *level* of this architecture. As Figure 4.11 shows, at level $\mathcal{L} = 1$ patches randomly sampled from $\{I_T^{(S_1)}\}$ are used to learn the first set of K context filters $C^{(1)} = [C_{11} | \dots | C_{K1}]^\top$, where C_{ij} is $p \times p$ pixels. This captures context information in the range of p pixels in each direction. Then, Gaussian smoothing with σ_{MRC} and downscaling by a factor 2 (i.e., image width and height are halved) and bi-cubic interpolation are applied to the tubularity maps $\{I_T^{(S_1)}\}$ to obtain $\{I_T^{(S_2)}\}$ and access the level $\mathcal{L} = 2$. Context filter learning is applied to these downsampled images to learn a second set of K filters $C^{(2)} = [C_{12} | \dots | C_{K2}]^\top$, C_{ij} is again $p \times p$ pixels. This process is repeated until level $\mathcal{L} = l_M$ is reached and its associate set of context filters $C^{(l_M)} = [C_{1l_M} | \dots | C_{Kl_M}]^\top$ is learned.

Notice, context filters size is the same (i.e., $p \times p$) at each level \mathcal{L} but images width and height are halved from $\mathcal{L} = i$ to $\mathcal{L} = i + 1$, doubling the context range in each direction. This makes this context learning solution *multi-range*. Moreover, although the number of available training patches reduce as images are downsampled, K -means can be still employed with good clustering performance, making the proposed solution fast and efficient.

4.4.2 Experiments and results

My target application is fully automated image-level tortuosity estimation, therefore I validate the proposed segmentation architecture on IVCN140, including 140 (384×384 pixel) corneal nerve images from healthy and unhealthy subjects showing a wide range of tortuosity characteristics (full description in Section 3.2.2).

Parameter setting. Table 4.2 shows the complete list of parameters I adopted or learned to train and test the proposed segmentation pipeline. Here, I provide explanations and

Table 4.2 Parameters of the proposed segmentation architecture.

System parametrisation

Parameter	Adopted value	Description
σ_1	{2,3,4}	SCIRD elongation.
σ_2	{2,3}	SCIRD width.
k	{0,0.025,..., 0.1}	SCIRD curvature.
θ	{15, 30, ..., 180}	SCIRD rotation angles.
α	1	SCIRD contrast enhancement.
$p \times p$	15×15 pixels	Context filters size.
ϵ_{ZCA}	0.001	Patch-level whitening in Spherical K -means [36].
K	100	Number of context filters learned for each pyramid level.
\mathcal{L}	3	Number of pyramid levels set manually in order to incorporate multi-range context.
σ_{MRC}	1	Sigma for Gaussian smoothing when learning multi-range context filters.
N_T	100	Number of trees in the random forest.
N_l	Learned	Maximum number of samples in each leaf of the random forest.

guidelines with the aim of facilitating adaptation to other data sets, potentially including different curvilinear structures.

I set SCIRD parameters according to the guidelines reported in Section 4.2.4 (same setting used for IVC100). Notice that the number of SCIRD free parameters and the experience required to set them is comparable with other HCFs (e.g., Gabor, Frangi, OOF).

I learned 100 15×15 filters using at least 100,000 patches for each context level. Increasing the number of context filters is expected to improve segmentation performance, at a price of a slower segmentation. Experiments suggest that 100 filters per level represent a good compromise between training/testing speed and tortuosity estimation performance. Moreover, I tested the sensitivity of the segmentation performance to the filter size and found that performance was practically unchanged when $p = \{11, 15, 21\}$ pixels. When setting ϵ_{ZCA} I considered the compromise between maximising filter sharpness and limiting noise amplification, as suggested by Coates and Ng [36]. I set σ_{MRC} and \mathcal{L} in the multi-range context architecture to capture short-, medium- and long-range context. In particular, σ_{MRC} controls the level of detail preserved in each context level, while \mathcal{L} controls the range of context to be captured (i.e. the number of context levels). According to Criminisi et al. [40], increasing N_T should lead to better classification performance, at a cost of slower predictions. I found experimentally that $N_T = 100$ is enough for my application. The depth of the forest is automatically learned using the out-of-bag error on the training set [22].

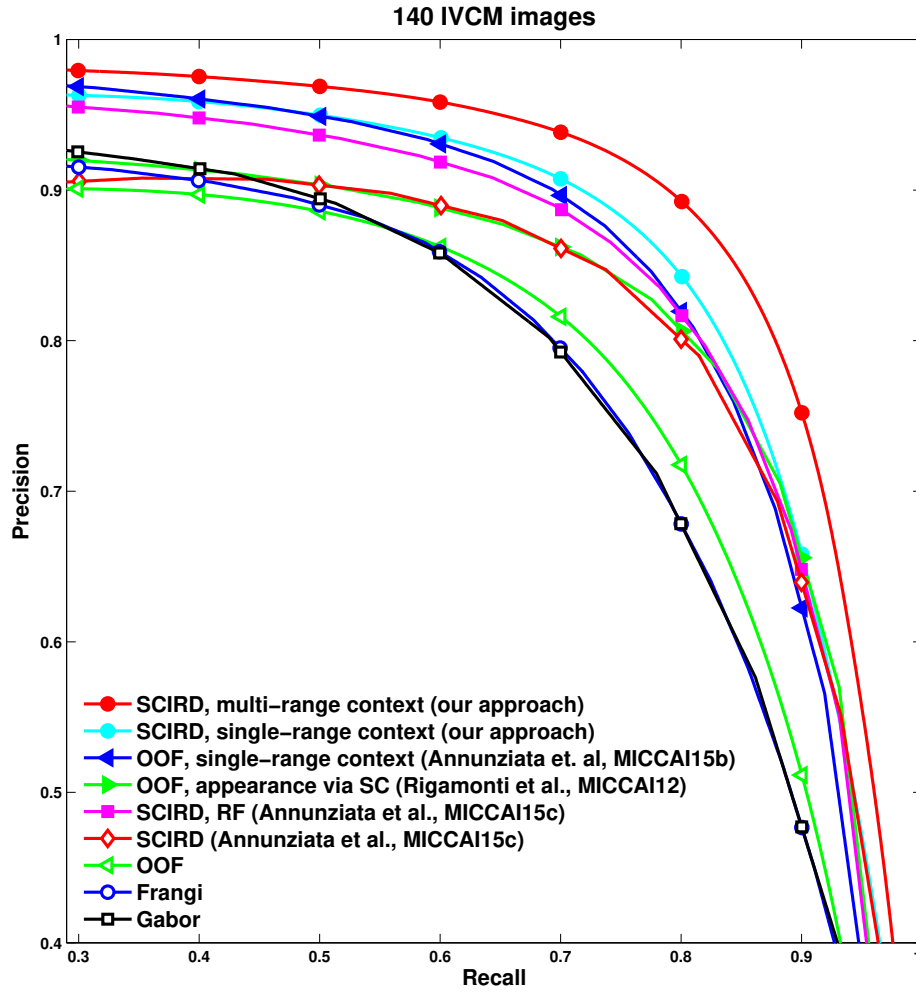


Figure 4.12 Precision-recall curves for the corneal nerve fibre centreline detection task (pixel-level classification). Each curve is obtained by averaging the results of 5 random cross-validation trials in which the whole data set, including 140 images, is randomly split in 2 equal partitions.

Results and discussion. I compare the proposed segmentation method with the recent hybrid solution proposed by Rigamonti and Lepetit [111] as well as with other baselines, including widely used and state-of-the-art ridge detectors: Frangi [49], Gabor filters [123] and Optimally Oriented Flux (OOF) [83]. Performance is reported in terms of PRC and is averaged over 5 random sub-sampling cross-validation runs, using 70 images for training and the rest for testing in each run (Figure 4.12). The comparison among HCFs shows that SCIRD performs best, therefore justifying the choice of adopting SCIRD as base HCF on top of which I learn multi-range context filters. Using SCIRD filter bank responses as input to a

RF, i.e. “SCIRD, RF”, leads to a significant improvement compared to plain SCIRD, especially at low-medium recall rates. It is worth noting that this result on IVC140 differs from the one on IVC100 (Section 4.2.4). However, IVC140 includes images from normal and herpetic patients in addition to all the images from dry eye patients in IVC100, therefore presenting a much wider spectrum of appearance characteristics compared to IVC100. In this setting, a supervised approach that combines the responses of multiple filters seems to be more suitable than plain SCIRD. Combining HCFs with learned (single-range) context filters leads to better segmentation results. However, using SCIRD instead OOF in the combination, yields better performance. A noticeable improvement is achieved by employing multi-range context filters, i.e. “SCIRD, multi-range context”, showing the benefit of learning a discriminative model capturing context at different levels.

Qualitative comparisons with the baseline methods, shown in Figure 4.13, suggest that improved performance is largely due to the multi-range context model which reduces considerably the amount of false positives far from target structures and improves the connectivity of corneal nerve fibres appearing fragmented. Nevertheless, the proposed segmentation pipeline is fast as it exploits context information efficiently: since a single discriminative model (i.e., a single RF) is employed, a whole corneal nerve image is segmented in about 30 seconds, using unoptimized MATLAB (R2014a) code on a machine equipped with Intel i7-4770 CPU at 3.4 GHz.

4.5 Conclusions

In this chapter I presented a hybrid approach to curvilinear structure segmentation. It consists of a novel class of HCFs (SCIRD) which are then leveraged by a learning-based architecture of multi-range context filters.

The proposed ridge detector (SCIRD) is based on curved-support Gaussian, and its formulation is such that it is simultaneously invariant to orientation, scale, and unlike its peers, curvature invariant. Experimental results show that SCIRD outperforms current state-of-

the-art HCFs on 3 challenging data sets, two of which were used in the recent literature for similar methods.

Boosting⁷ HCFs with learned filters has recently emerged as a successful technique to compensate for limits of HCFs and DLAs. I have proposed a novel combination method in which HCFs are paired with learned context filters to enhance pixel representation including inter-object relationships. Quantitative results suggest that the proposed approach outperforms a previous combination method. Moreover, it can be used for different image modalities and can get top-rank performance running in a few minutes only.

Finally, I have discussed the limitations of single-range context filters and proposed a solution to make them multi-range. Experimental results with 140 IVCN images show that combining SCIRD with multi-range context filters performs favourably with respect to state-of-the-art segmentation methods based on single discriminative models, without increasing the computational cost (i.e. using a single classifier).

This solution represents the segmentation module of the fully automated tortuosity estimation system discussed herein. In the next chapter, I will introduce and validate the tortuosity estimation module and discuss the impact of the segmentation on the tortuosity estimation results.

⁷Combining.

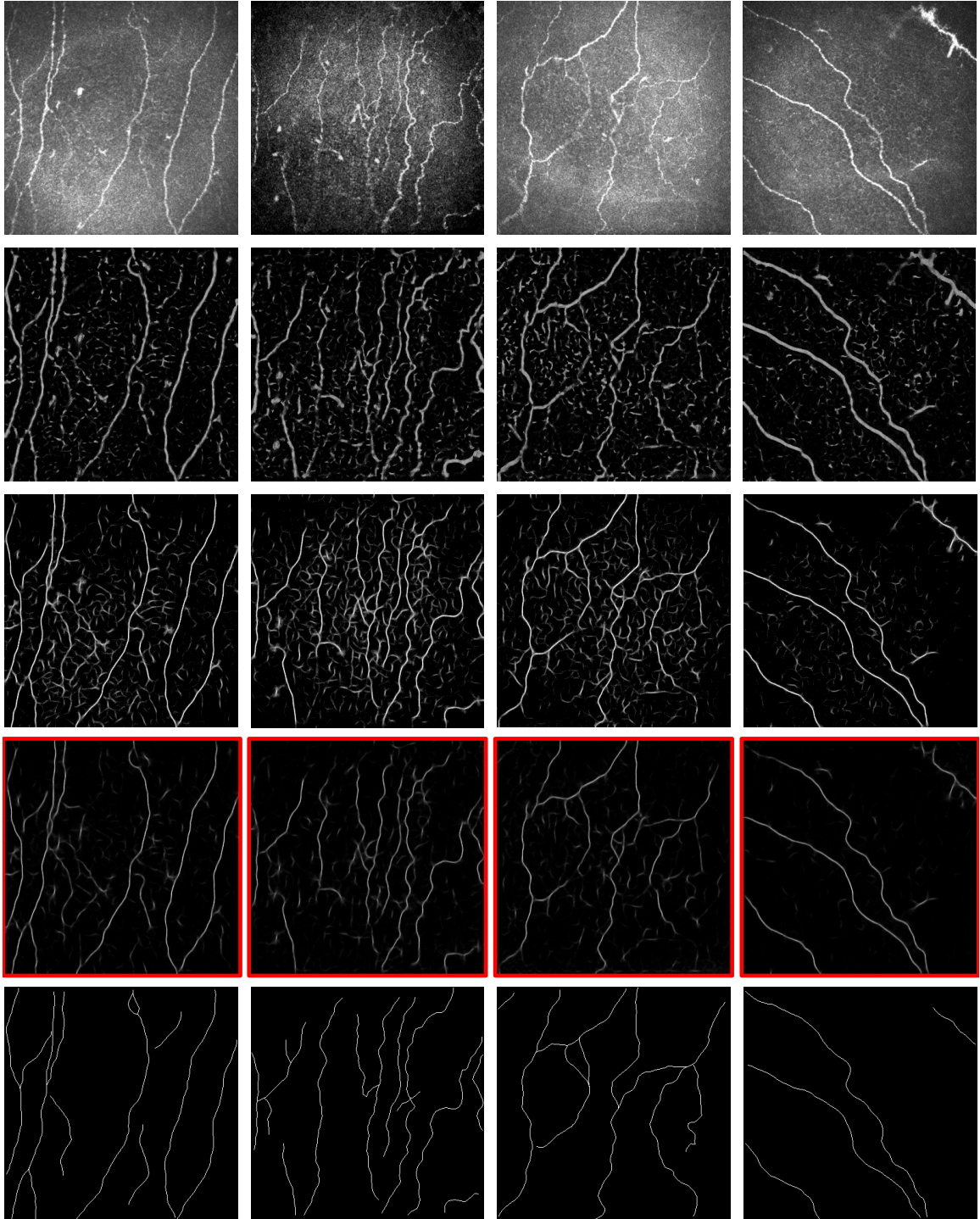


Figure 4.13 Examples from segmentation experiments on IVCM140. First row: original images. From second to fourth row: probability maps obtained by combining OOF with learned *appearance* filters [111], combining OOF with *single-range* context filters and combining SCIRD with *multi-range* learned context filters (proposed approach). Last row: ground truth.

Chapter 5

Tortuosity Estimation

5.1 Introduction

This chapter describes the approach I adopted to address the limitations of state-of-the-art tortuosity estimation systems outlined in Section 2.4.2. In particular, Section 5.2 describes a new paradigm to tortuosity estimation based on machine learning. The latter includes a richer representation of tortuosity (both at structure and image level), based on multi-scale features (Section 5.2.1 and 5.2.2), and the methodology to identify the most discriminative tortuosity features (and their combination) among the pool investigated (Section 5.2.3). The validation is presented and discussed in Section 5.2.4. Finally, Section 5.3 describes a new representation tool, the *tortuosity plane*, adopted to leverage tortuosity interpretation.

5.2 A machine learning approach to tortuosity estimation

Tortuosity estimation is a difficult task as different anatomical structures in the context of different pathologies can show different tortuosity characteristics. Most of the methods reported so far lack adaptability as they rely on fixed, *postulated* combinations of features to characterise tortuosity. Moreover, different frequencies of direction changes are ignored, while an in-depth visual investigation of corneal images, for instance, suggests that

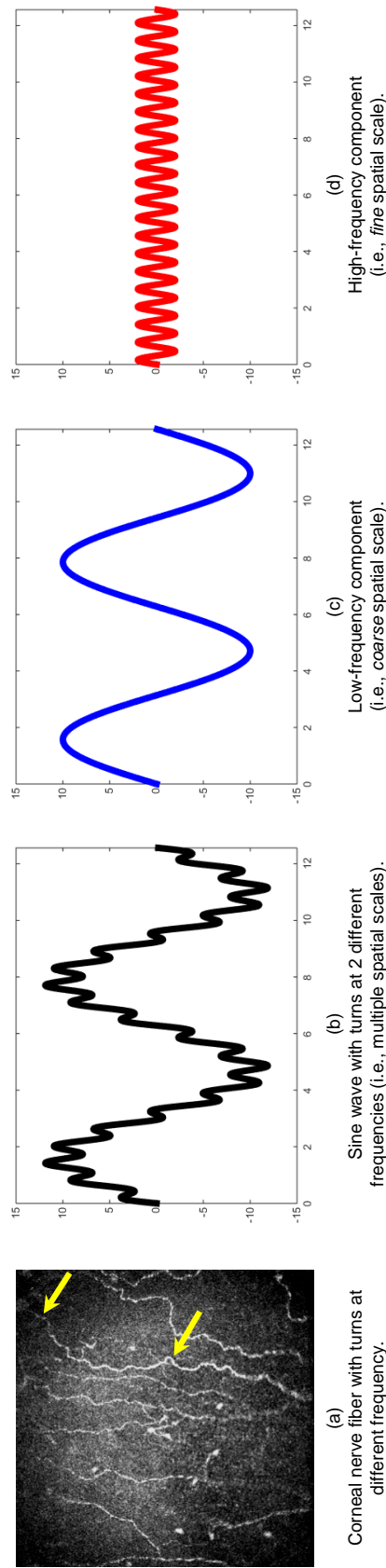


Figure 5.1 Example of multiple spatial scales observed in a corneal nerve fibre (a, yellow arrows). Multi-scale decomposition (c,d) of a simple model of corneal nerve fibre obtained as sum of two sine waves with different frequencies (b).

a multiscale approach for tortuosity features extraction is indeed needed (Figure 5.1). Finally, curvature estimation algorithms are typically based on finite differences leading to noisy estimates, especially when segmentations are obtained automatically.

The method proposed here addresses all the issues above. It can be summarised in four steps (Figure 5.2):

1. compute multi-scale shape features for each curvilinear structure to obtain its tortuosity representation;
2. combine shape features computed from each curvilinear structure to obtain image-level features;
3. use feature selection (FS) to identify the most discriminative set of image-level features, which become the tortuosity representation *for the specific application* considered;
4. use the set of features identified to train a regressor assigning new images to a tortuosity grade on a fixed scale.

5.2.1 Multi-scale tortuosity representation of a single curvilinear structure

Curvature-based measures tend to outperform the combination of distance measure (DM, i.e. maximum deviation from chords) and number of inflection points, when high sampling rates are used to represent vessel centrelines ([93]; see also Section 2.3).

We showed in [5] that multi-scale curvature-based measures, if estimated accurately, can lead to accuracy comparable or even higher than that of two experienced observers, when the third is taken as reference. Our algorithm for curvature estimation was based on a multi-window ellipse/line fitting approach which can be time consuming, although sub-sampling could achieve a reasonable trade-off between performance and time.

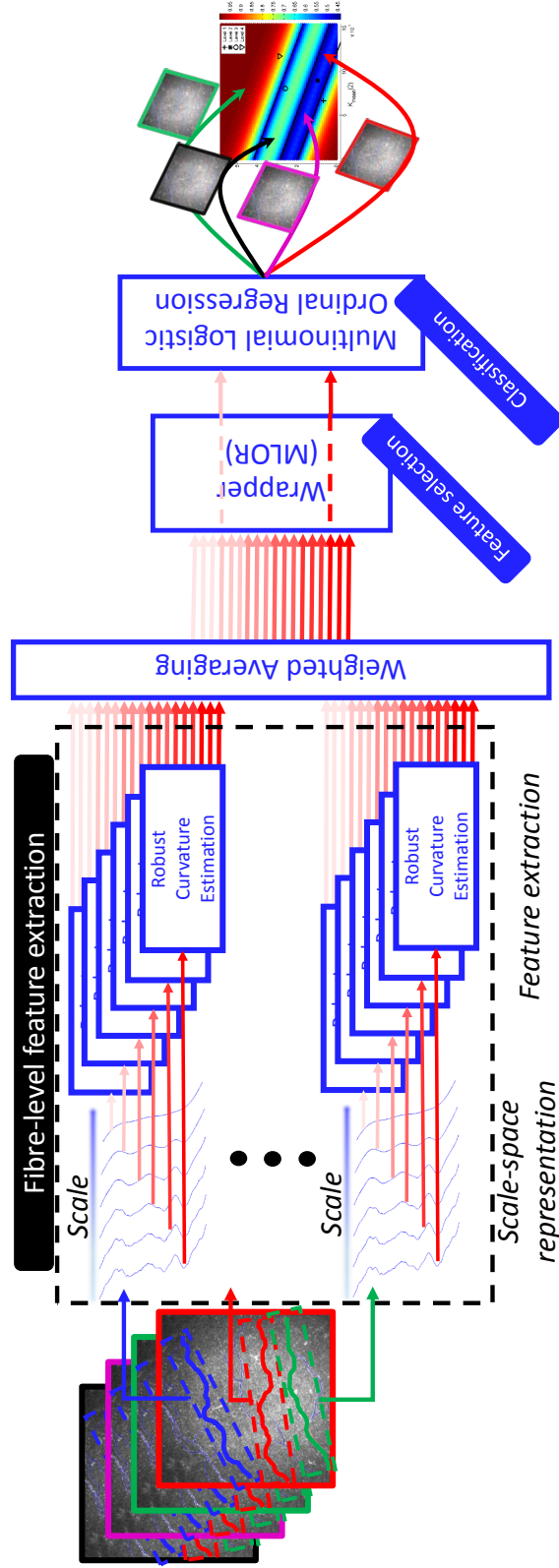


Figure 5.2 Block diagram of our tortuosity estimation framework: wrapper-based feature selection followed by multinomial logistic ordinal regression. First, for every corneal nerve fibre segment detected in an image, robust spline fitting is applied for fast and accurate curvature estimation at each level of the fibre's scale-space representation. Mean curvature $k_{mean}(t)$, twistedness $d_{ip}(t)$, and maximum curvature $k_M(t)$ are computed for each scale $t \in \{1, \dots, t_M\}$ and the related feature vector v_f is built. Second, weighted (fibre length is used as weight) averaging across all the fibres creates the pool of image-level features v_{img} . Third, a wrapper-based FS technique is employed for identifying the most discriminative combination of features and their scale. Finally, a multinomial logistic ordinal regressor is used to assign image-level tortuosity.

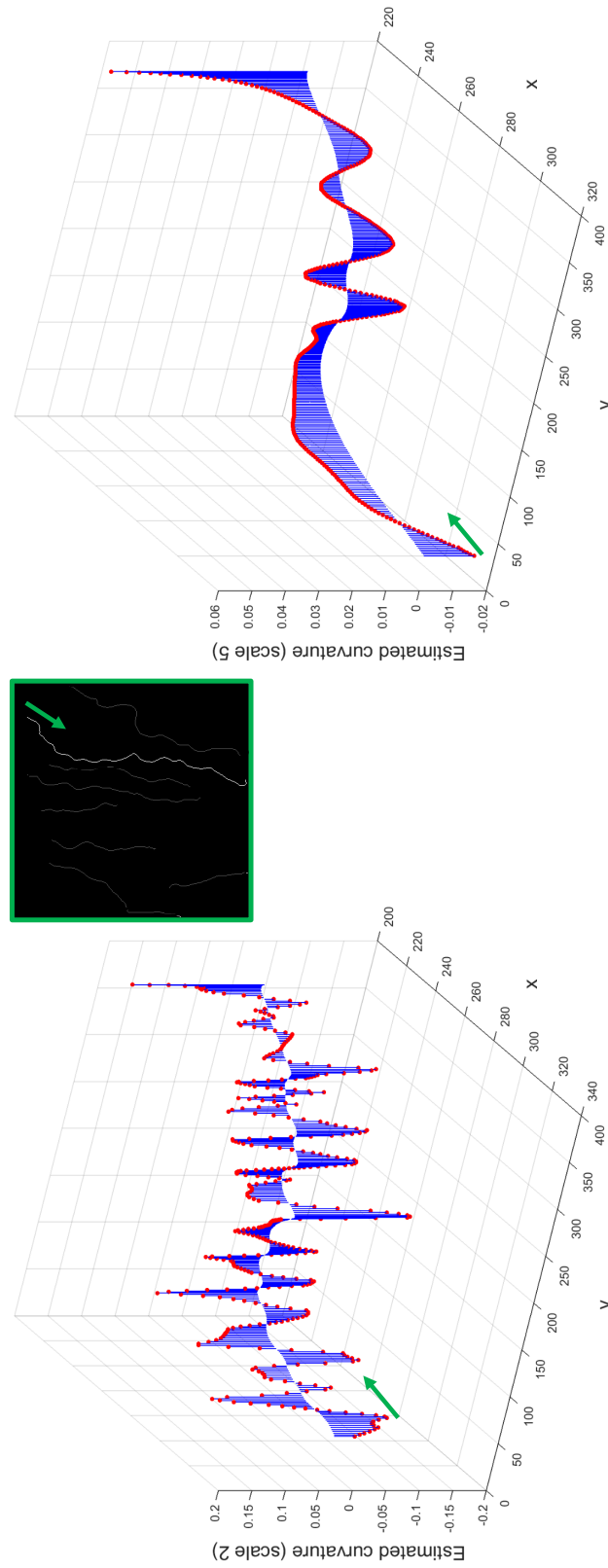


Figure 5.3 Accurate multi-scale curvature estimation via cubic spline fitting. Stem plots illustrate the estimated local curvature for the automatically segmented corneal nerve fibre shown in the top image. Left, right: estimated curvature at scale 2 and 5, respectively. The green arrows indicate starting point ($l = 0$) and direction ($l > 0$).

Here, I use spline fitting to replace the multi-window ellipse/line fitting for fast and accurate curvature estimation. The structures of interest are normally rotated first or represented in parametric coordinates. I choose the latter to avoid the additional computational burden and potential inaccuracies introduced by rotations. To counteract the noise introduced by the automated centreline segmentation, I adopt a robust version of cubic spline fitting, which reduces the influence of outliers¹. This yields accurate curvature estimates, supporting reliable estimation of inflection points.

Derivatives are computed locally and *analytically* from the parameters of the fitted splines. Changes in curvature sign along the curves are the estimated inflection points (Figure 5.3).

In summary, the following shape features are extracted from each fibre [5]:

- the average curvature along the fibre, k_{mean} ;
- the “twistedness”, or density of inflection points, d_{ip} ;
- the maximum curvature along the fibre, k_M .

I propose to represent curvilinear structures for tortuosity estimation using a scale-space representation, taking into account the different frequency of direction changes, as shown in Figure 5.4. The multi-scale version of the features listed above becomes: $\{k_{mean}(t)\}$, $\{d_{ip}(t)\}$ and $\{k_M(t)\}$, for $t \in \{1, \dots, t_M\}$, where t_M is the coarsest resolution used.

5.2.2 Image-level tortuosity features for image classification

Our next task is *to assign a tortuosity grade to a whole image*. This is done, for instance, in the clinical assessment of images of the corneal nerves. The task is non-trivial as images contain variable numbers of corneal nerve fibres of varying lengths, which could show considerably different tortuosity characteristics. I turn the features computed for individual fibres into *image-level features* by computing, at each scale t and for each feature, a weighted average over all fibers in the image, where the weights are the fibre lengths. The

¹MATLAB code available at <http://www.mathworks.com/matlabcentral/fileexchange/13812-splinefit>

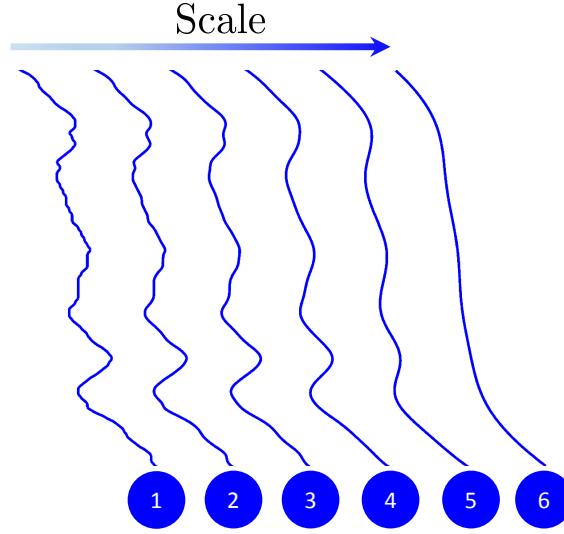


Figure 5.4 Multi-scale analysis of a corneal nerve fibre. From left to right: original fibre including turns at all frequencies (spatial scale 1), and its smoothed versions at spatial scales from 2 to 6 to take into account turns at intermediate and high frequencies.

weights model the observation that longer fibres are more informative, overall, than shorter ones.

Denoting the length of the i -th fibre at scale t as $l_i(t)$, and the total number of fibres within an image as N_f , our image-level features are defined as follows:

- $K_{mean}(t) = \frac{\sum_{i=1}^{N_f} l_i(t) k_{mean}(t)}{\sum_{i=1}^{N_f} l_i(t)}$;
- $D_{ip}(t) = \frac{\sum_{i=1}^{N_f} l_i(t) d_{ip}(t)}{\sum_{i=1}^{N_f} l_i(t)}$;
- $K_M(t) = \frac{\sum_{i=1}^{N_f} l_i(t) k_M(t)}{\sum_{i=1}^{N_f} l_i(t)}$.

5.2.3 Feature selection and tortuosity prediction

FS identifies the most discriminative features and the most important spatial scales for the tortuosity assessment. This choice makes the tortuosity estimation framework highly adaptable.

Of the several FS strategies proposed [56, 85, 108], I adopt a wrapper-based strategy in which a learning algorithm is used to score repeatedly subsets of features according to their

predictive power. Exhaustive search can be employed as our feature pool is relatively small (i.e. less than 20, including different spatial scales), guaranteeing that the search returns the true absolute extremum. I use a multinomial logistic ordinal regressor (henceforth, MLOR), a low-complexity, hence fast multi-class classifier taking into account the order of tortuosity grades (1 to 4, i.e., ranking), modelled by a *logit* link function. Specifically, MLOR models the *log cumulative odds* [18], the logarithm of the ratio of the probability of a class preceding or equal to class j in the class ranked list, $P(y_c \leq c_j)$ and the probability of the same class following class j , $P(y_c > c_j)$. To limit model complexity and make FS more efficient, we assume that the effects of features f_i are the same for all classes on the logarithmic scale. This means that the model has different intercepts (γ), but common slopes (β) among classes (proportional odds assumption). Thus, the MLOR model is

$$\ln \left(\frac{P(y_c \leq c_j)}{P(y_c > c_j)} \right) = \gamma_j + \beta_1 f_1 + \beta_2 f_2 + \cdots + \beta_s f_s \quad (5.1)$$

for $j \in \{1, 2, 3\}$, where s is the feature space cardinality.

Once the most discriminative combination of features is selected by exhaustive search, these are extracted from training images and a new MLOR model is learned and used to predict unseen images.

5.2.4 Experiments and results

I compute image-level tortuosity measures (mean, max and density of inflection points) in a scale-space representation of 6 spatial scales (i.e. $t_M = 6$)², therefore assessing the discriminative power of 18 features in total.

I compare the performance of our multi-scale tortuosity estimation approach with single-scale methods reported to perform very well in comparative tests: tortuosity density (TD, [54]), distance measure (DM, [64]), slope chain coding (SCC, [23]), τ_5 [61]. Moreover, I report the results obtained using a tortuosity estimation algorithm based on multi-window curvature estimation [5].

²I set the maximum spatial scale $t_M = 6$ as, for $t_M > 6$, almost all the frequency components of the corneal nerve fibres in IVC140 are smoothed out.

Since the IVC140 data set was annotated by three experienced observers independently (clinical collaborators at Harvard Medical School, reported as Obs_1 , Obs_2 , Obs_3), I take the classification by each observer in turn as ground truth, and compare the performance of ours and aforementioned methods with that of the other two observers. Therefore, MLOR coefficients are learned using a leave-one-out cross-validation for each observer taken as reference. Notice that 11 images, corresponding to less than 8% of the entire data set, were excluded (only for the tortuosity estimation performance evaluation, not for the segmentation) due to complete disagreement among the observers (i.e. all the observers assigned a different tortuosity level to those images), confirming the difficulty of tortuosity modelling. There are mainly two reasons behind this choice. First, given that I compare the performance of automated algorithms (trained on one observer) with that of the other two observers, excluding those 11 images is indeed a more challenging scenario in terms of multi-class classification performance. In fact, for those 11 images the tortuosity estimation algorithm would perform at least as well as the other two observers. Second, when FS is applied to identify the best combination of tortuosity features for the entire data set, I adopt the majority voting criterion to establish the ground truth tortuosity level of each image, therefore leading to ambiguity on those 11 images. Other choices and assumptions could be made (e.g. averaging the ground truth tortuosity level as assigned by each observer for each image), and those images could be included (as they would still carry useful information), but I reckoned that discarding this small amount of images was a reasonable choice for this pilot-level study.

I investigate the impact of each building block on the tortuosity estimation performance of the proposed framework for fully automated tortuosity estimation. Specifically, (1) I compare the performance of the proposed tortuosity estimation algorithm with state-of-the-art methods, at a parity of segmentation; (2) I assess the performance when replacing a conventional segmentation method (i.e. based on the “locally-straight” assumption) with the proposed segmentation approach, at a parity of tortuosity estimation algorithm.

Table 5.1 Tortuosity estimation: comparison of tortuosity estimation algorithms at a parity of segmentation approach. For tortuosity ground truth, each experienced observer is taken as reference in turn (Obs_1 , Obs_2 and Obs_3). The segmentation approach used for these experiments is our “*SCIRD, multi-range context*” shown to outperform others methods on the IVC140 data set (see Figure 4.12). The first two columns in each table show the performance of the others observers against the one used as reference. The other columns in each table report several tortuosity estimation algorithms. Specifically, *MSMW* is the algorithm based on multi-scale-multi-window tortuosity [5], *MSSPLINE* is the proposed approach based on multi-scale rotation invariant spline fitting (results in boldface), *MSSPLINE (no FS)* is the latter approach without using tortuosity feature selection, TD is the tortuosity density index [54], *DM* is the distance measure [64], *SCC* is the tortuosity estimation algorithm based on slope chain coding [23] and τ_5 is the normalised integral of the squared curvature [61].

Ground Truth = Obs_1									
Performance measure	Obs_2	Obs_3	MSMW	MSSPLINE	MSSPLINE (no FS)	TD	DM	SCC	τ_5
Acc	0.7678	0.7451	0.7080	0.7408	0.7066	0.6247	0.6236	0.6194	0.6303
Se	0.5736	0.5194	0.4729	0.5194	0.4651	0.3333	0.3411	0.3333	0.2171
Sp	0.8457	0.8219	0.7825	0.8113	0.7904	0.6797	0.6732	0.6704	0.8083
PPv	0.5729	0.5547	0.4815	0.5184	0.4630	0.1165	0.3842	0.1129	0.4864
Npv	0.8329	0.8223	0.7949	0.8204	0.7945	0.8379	0.8410	0.8388	0.7346
MSE	0.4729	0.5736	0.6202	0.5504	0.6512	1.1705	1.0543	1.0853	2.6977
MAE	0.4419	0.5116	0.5581	0.5039	0.5736	0.8295	0.7907	0.8062	1.3333

Ground Truth = Obs_2									
Performance measure	Obs_1	Obs_3	MSMW	MSSPLINE	MSSPLINE (no FS)	TD	DM	SCC	τ_5
Acc	0.7791	0.7513	0.7619	0.7838	0.7802	0.6304	0.6301	0.6260	0.6661
Se	0.5736	0.5271	0.5426	0.5814	0.5736	0.3101	0.3101	0.3023	0.3101
Sp	0.8455	0.8254	0.8334	0.8478	0.8506	0.7055	0.7024	0.6997	0.8044
PPv	0.5977	0.5580	0.5473	0.5807	0.5678	0.2278	0.3564	0.0928	0.5479
Npv	0.8507	0.8317	0.8386	0.8560	0.8514	0.7864	0.8370	0.8350	0.7774
MSE	0.4729	0.5659	0.6202	0.5349	0.5426	1.2946	1.2946	1.3411	2.6124
MAE	0.4419	0.5039	0.5116	0.4574	0.4651	0.8915	0.8915	0.9070	1.2326

Ground Truth = Obs_3									
Performance measure	Obs_2	Obs_1	MSMW	MSSPLINE	MSSPLINE (no FS)	TD	DM	SCC	τ_5
Acc	0.7329	0.7381	0.6735	0.7266	0.6777	0.6366	0.6411	0.6388	0.6232
Se	0.5271	0.5194	0.4419	0.5271	0.4419	0.4109	0.4186	0.4109	0.2868
Sp	0.8290	0.8275	0.6994	0.7409	0.7299	0.6008	0.5983	0.5969	0.7848
PPv	0.5495	0.5459	0.4223	0.5057	0.4293	0.2892	0.3187	0.1715	0.3269
Npv	0.7930	0.8001	0.7711	0.8286	0.7617	0.7405	0.8778	0.8766	0.7306
MSE	0.5659	0.5736	0.7907	0.6124	0.7442	1.1008	1.1163	1.1473	2.2171
MAE	0.5039	0.5116	0.6357	0.5194	0.6202	0.7597	0.7597	0.7752	1.1473

I carried out FS on the pool of multi-scale features using the majority voting ground truth for tortuosity estimation as reference³. I found through a 20-fold nested cross-validation (training, validation, test sets) on IVC100 that the best combination of image-level tor-

³The majority voting ground truth assumes that the true tortuosity level of each image is the one assigned by the majority of experienced observers, independently.

tuosity features was $\{K_{mean}(2), K_{mean}(5)\}$ in 17 out of 20 cases[5], suggesting that (1) a multi-scale approach is more suitable to tortuosity estimation than single-scale indices reported in literature, and (2) the feature selection procedure is very stable. I adopted this combination of features in these experiments.

Comparison of tortuosity estimation algorithms at a parity of segmentation. Table 5.1 shows the experimental evaluation for the tortuosity estimation task when state-of-the-art tortuosity estimation approaches are used together with the best segmentation approach (i.e., “SCIRD, multi-range context”, as shown in Figure 4.12). First, experimental results adopting each expert observer as reference in turn, suggest that the proposed tortuosity estimation method based on multi-scale cubic splines (MSSPLINE) outperforms the multi-scale-multi-window approach (MSMW) previously proposed [5]. Moreover, MSSPLINE is much faster than MSMW as shown in Table 5.5, thus successfully addressing the main drawback of MSMW. Second, MSSPLINE performs significantly better than the baselines for all the observers used as reference, thus suggesting that a multi-dimensional tortuosity representation feeding a MLOR model is more suitable for this task. Third, tortuosity FS yields a substantial extra performance when Obs_1 and Obs_3 are used as reference. This can be justified by the fact that some features may be highly correlated, noisy due to segmentation imperfections or not discriminative enough. Notice that FS does not improve tortuosity estimation performance when Obs_2 is used as reference. However, matching the performance obtained with a larger feature vector is still an important advantage in terms of computational efficiency (i.e., less features to be computed at run time). Finally, the average performance of the proposed fully automated tortuosity quantification system (i.e., “SCIRD, multi-scale context + MSSPLINE”, reported in this table as MSSPLINE) matches or even exceeds (e.g. by about 3% in terms of accuracy, when Obs_2 is used as ground truth) the tortuosity estimation performance of at least one of the other two expert observers, when compared against the one taken as reference.

Comparison of segmentation algorithms at a parity of tortuosity quantification approach. Table 5.2 shows the tortuosity estimation performance when replacing a conventional segmentation method [111] with the proposed one. The tortuosity estimation al-

Table 5.2 Tortuosity estimation: comparison of segmentation algorithms at a parity of tortuosity quantification approach. For tortuosity ground truth, each experienced observer is taken as reference in turn (Obs_1 , Obs_2 and Obs_3). The tortuosity quantification approach used for these experiments is our *MSSPLINE* shown to outperform other methods on our data set (see Table 5.1). The first two columns in each table show the performance of the others observers against the one used as reference. The other columns in each table report tortuosity quantification performance when the following approaches are used: manual segmentation (*Manual*); the proposed “*SCIRD, multi-range context*” segmentation algorithm (*Proposed*, in bold face), based on curved-support and multi-range context filters; the segmentation method proposed by [111], based on “locally-straight” and appearance filters.

Ground Truth = Obs_1					
Performance measure	Obs_2	Obs_3	Manual	Proposed	Rigamonti et al.
Acc	0.7678	0.7451	0.7448	0.7408	0.7182
Se	0.5736	0.5194	0.5426	0.5194	0.4806
Sp	0.8457	0.8219	0.8068	0.8113	0.7888
PPv	0.5729	0.5547	0.5590	0.5184	0.4745
Npv	0.8329	0.8223	0.8240	0.8204	0.8064
MSE	0.4729	0.5736	0.5736	0.5504	0.6589
MAE	0.4419	0.5116	0.4961	0.5039	0.5659

Ground Truth = Obs_2					
Performance measure	Obs_1	Obs_3	Manual	Proposed	Rigamonti et al.
Acc	0.7791	0.7513	0.8039	0.7838	0.7327
Se	0.5736	0.5271	0.6202	0.5814	0.4806
Sp	0.8455	0.8254	0.8618	0.8478	0.8111
PPv	0.5977	0.5580	0.6262	0.5807	0.4715
Npv	0.8507	0.8317	0.8696	0.8560	0.8217
MSE	0.4729	0.5659	0.4729	0.5349	0.7829
MAE	0.4419	0.5039	0.4109	0.4574	0.5969

Ground Truth = Obs_3					
Performance measure	Obs_2	Obs_1	Manual	Proposed	Rigamonti et al.
Acc	0.7329	0.7381	0.6828	0.7266	0.6589
Se	0.5271	0.5194	0.4496	0.5271	0.4031
Sp	0.8290	0.8275	0.7055	0.7409	0.6811
PPv	0.5495	0.5459	0.4025	0.5057	0.2946
Npv	0.7930	0.8001	0.7858	0.8286	0.7737
MSE	0.5659	0.5736	0.6899	0.6124	0.8217
MAE	0.5039	0.5116	0.5969	0.5194	0.6667

gorithm used here is our *MSSPLINE*, shown to outperform other methods in Table 5.1. Experimental results suggest that modelling tortuous structures through *SCIRD* and using

Table 5.3 Confusion matrices for the tortuosity estimation task obtained using the proposed method. For tortuosity ground truth, individual observers are used in turn (Obs_1 , Obs_2 and Obs_3).

		Estimated level						Estimated level						Estimated level			
		1	2	3	4			1	2	3	4			1	2	3	4
Obs₁ level	1	22	12	1	0	Obs₂ level	1	23	9	2	0	Obs₃ level	1	1	16	2	0
	2	12	26	4	1		2	10	25	3	1		2	1	45	5	2
	3	0	14	12	8		3	0	13	10	7		3	0	20	8	6
	4	0	1	9	7		4	0	2	7	17		4	0	2	7	14

Table 5.4 Cohen’s kappa and modified kappa for the tortuosity estimation task. For ground truth, individual observers are taken as reference in turn (Obs_1 , Obs_2 and Obs_3). We compare the performance of the proposed method (i.e., *SCIRD*, *multi-range context* + *MSSPLINE*, indicated as *Ours* here) with the expert observers, in terms of Cohen’s kappa (**K**, first row) and modified kappa (**K/K_M**, second row).

	GT = Obs ₁			GT = Obs ₂			GT = Obs ₃		
	Obs ₂	Obs ₃	Ours	Obs ₁	Obs ₃	Ours	Obs ₁	Obs ₂	Ours
K	0.42	0.34	0.33	0.42	0.36	0.43	0.34	0.36	0.28
K/K_M	0.47	0.41	0.37	0.47	0.44	0.48	0.41	0.44	0.43

Table 5.5 Comparison in terms of running time for each tortuosity estimation algorithm alone and in combination with the automated segmentation algorithm proposed herein. The proposed spline-based curvature estimation, i.e. *MSSPLINE* (results in boldface), is significantly faster than our previous multi-window solution, i.e. *MSMW*. Experiments were carried out on Intel i7-4770 CPU @ 3.4 GHz, using MATLAB code. Each image was 384×384 pixels.

Execution Time (140 images)	MSMW	MSSPLINE	TD	DM	SCC	τ_5
Tortuosity (s)	7590	23.1	47.1	43	40.1	39
Segm. + Tortuosity (min)	196.5	70.4	70.8	70.7	70.7	70.7

multi-range context filters leads to a considerable improvement over conventional segmentation methods. This improvement is particularly remarkable as, for the first time to the best of my knowledge, it closes the gap between automated and manual segmentation for the tortuosity estimation task.

Confusion matrices and Cohen’s kappa. I further assess the performance of the proposed fully automated tortuosity quantification system by reporting confusion matrices and Cohen’s kappa values when each of the three observers is used in turn as reference. As Table 5.3 shows, most of the images classified incorrectly are classified in the immediately

following or preceding tortuosity level, a desirable property of our system, since tortuosity represents an ordinal variable. Results in terms of Cohen’s kappa and modified kappa (i.e. \mathbf{K}/\mathbf{K}_M , where \mathbf{K}_M is the maximum possible \mathbf{K} given marginal frequencies [44]) are reported in Table 5.4. Notice that given the different marginal frequencies for each tortuosity level (whose values change depending on the observer used as ground truth), the modified kappa represents a fairer index to assess agreement [44]. Nevertheless, I report the original kappa values for completeness. In both cases, the proposed system matches the performance obtained by the other expert observers, when compared to the one taken as reference.

Running time. I report running time for the tortuosity estimation algorithm alone and used in combination with automated segmentation in Table 5.5: about 70 minutes are required to estimate image-level tortuosity for 140 images (fully automated). Thank to the adopted spline-based curvature estimation, the running time is reduced considerably. The current parameter setting seems to represent a good compromise between speed and tortuosity estimation accuracy.

Applying the system to other tortuosity quantification problems. The proposed system was designed and developed to work on tortuous structures in general, and could potentially be applied to structures other than corneal nerve fibres. Specifically, SCIRD and the single-range context model were found particularly suitable for other curvilinear structures as well, e.g. neurites acquired with different modalities (Section 4.2.4 and 4.3.4), therefore presenting different characteristics in terms of contrast, resolution and context. The multi-range context model introduced here improved performance significantly on corneal nerve fibres; such improvements could be observed in similar scenarios, presenting non-target structures with same appearance characteristics or low signal-to-noise ratio. The approach for automated tortuosity quantification is highly versatile, as it is capable of identifying the most discriminative tortuosity features which may vary for different structures and, more importantly, for different pathologies. I maintained the segmentation and tortuosity estimation modules separate to allow an easier and more effective interpretation of the results, compared to solutions based on direct tortuosity feature extraction and image classification. In my experience, the main source of errors in tortuosity estimation is the segmentation

module, but training our hybrid segmentation solution requires a very limited amount of images and annotations, compared to fully-learned architectures. These choices should allow easier adaptation within the medical imaging domain, in which image data and annotations are not always abundant.

5.3 Tortuosity plane and confidence of the estimated tortuosity level

Previously proposed methods provide typically a single tortuosity level (or coefficient value), obfuscating the different effects of the factors considered, and limiting interpretations which could be important to the ophthalmologist for diagnosis.

Using the geometrical interpretation of the MLOR model, we can map each IVCN image onto a plane (*tortuosity plane* or TP) whose axes are the best tortuosity measures identified automatically by the feature selection procedure (in our case, weighted mean curvatures at spatial scales 2 and 5). Geometrically, the best weights γ and β of the MLOR model, define the best (in the MLOR sense) linear decision boundaries separating the plane into 4 regions corresponding to the four tortuosity levels (black lines in Figure 5.5). Any new corneal nerve image can be plotted as a point on the TP by computing the values of the two best tortuosity measures. The region containing the new point gives the estimated tortuosity level for the corneal nerve image.

Importantly, the TP provides also a level of confidence for the estimated tortuosity level, quantifying the reliability of the system. This confidence is given by the probability of belonging to one of the four regions (i.e., tortuosity levels) estimated automatically by the MLOR model, and is intuitively proportional to the distance of the point (i.e., image) from the linear decision boundaries. In fact, the closer the point to the boundary between two adjacent regions, the less reliable is the estimated tortuosity level. The level of confidence for all the points on the TP can be estimated once the system is trained (i.e. before analyzing the target images) and color-coded for immediate, intuitive visualization.

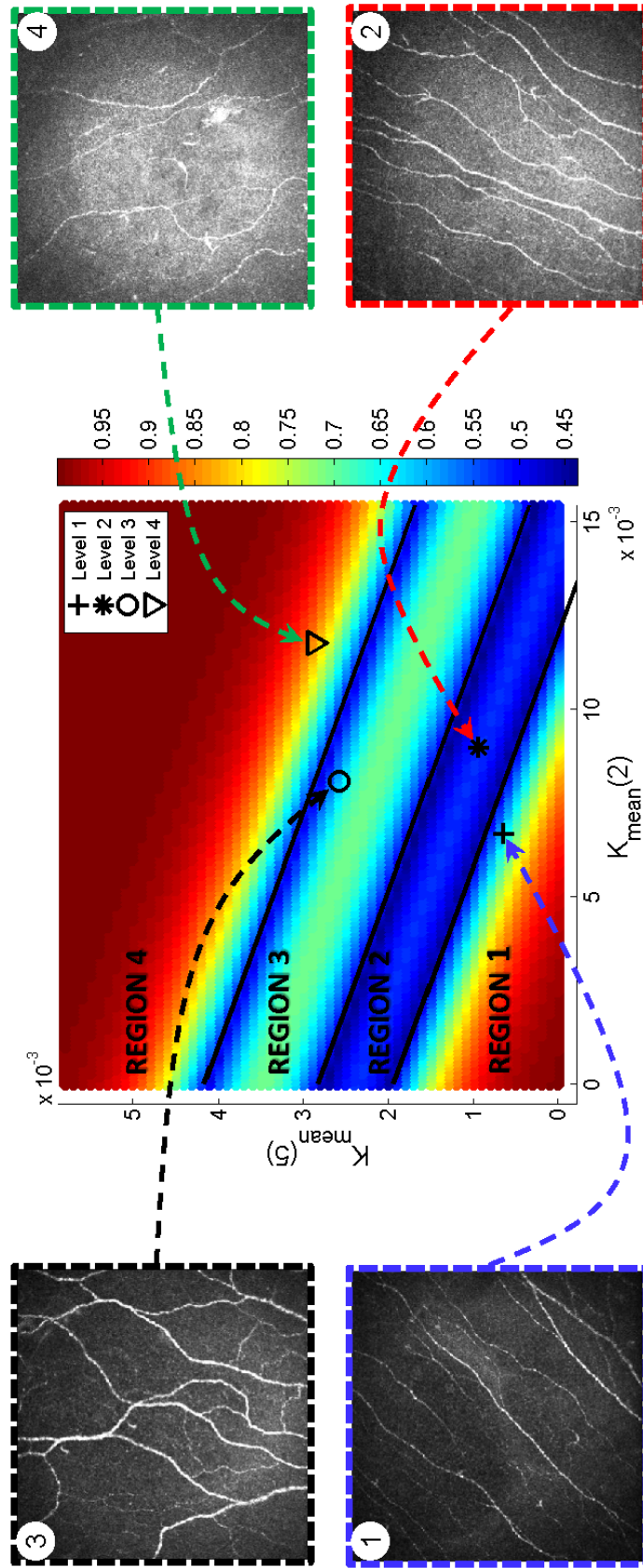


Figure 5.5 Tortuosity estimated for 4 corneal nerve images in IVCMI40 by the proposed method. Images are projected onto the tortuosity plane as points (markers indicate the tortuosity level assigned by the expert observers) whose coordinates are the estimated mean curvature at scale 2 (high frequency turns) and scale 5 (low frequency turns). The level of confidence for each tortuosity estimate is encoded with colour (red and blue mean high and low confidence, respectively). Intuitively, the confidence is related to the distance from the decision boundaries (indicated in black) separating each tortuosity region.

Figure 5.5 shows the TP obtained with the proposed method applied to a subset of IVCMI40 images during a random 20-fold cross-validation procedure. The 4 regions related to the 4 tortuosity levels are identified automatically and the best separating lines (visualized in black) traced. Then, the two best tortuosity measures are computed on each of the testing images and they are mapped onto the TP. Since each image belongs to a different tortuosity level, they are (correctly) mapped to four different regions. Visualizing the images as points on the TP allows an immediate and consistent interpretation.

First, image 1 is close to the boundary 1-2, so the confidence of tortuosity level 1 (instead of 2) is limited; this is consistent, as some fibres appear rather tortuous for level 1. The same considerations apply to image 3.

Second, although image 2 belongs to the tortuosity level with lowest overall accuracy (level 2), it is mapped to the center of the corresponding TP region, and therefore the confidence of the prediction is relatively high for level 2. Image 4 is classified with the highest confidence as it clearly shows an abrupt large change (scale 5) in the direction of the second fiber (counting from left to right), and also high-frequency changes in other fibres (scale 2). Previous, single-scale methods would not tease out this difference.

Third, the contribution of high and low frequency turns (i.e. curvatures at scales 2 and 5) for the predicted tortuosity level can be easily deduced. For instance, images 1 and 2 show a relatively low mean curvature at scale 5, but this is higher for images 3 and 4. Then, although image 2 is correctly assigned to tortuosity level 2, a higher mean curvature at scale 2 than image 3 is present: this suggests that, in our data set, the mean curvature at scale 2 (i.e. high frequency turns) is slightly less discriminative than the one at scale 5.

I ran a leave-one-out cross-validation simulating the scenario in which the tortuosity of a new unseen image is assessed. For each validation, the best boundaries were estimated automatically and the tortuosity level was assigned. Figure 5.6 shows the result of this process. I use different markers to indicate the true tortuosity level of each image as per majority voting ground truth. The four clusters reflecting the tortuosity levels are clearly visible and separated automatically by the proposed system, although some mis-classifications are present. Notice that the maximum error made by the system is always within 1 tortuosity

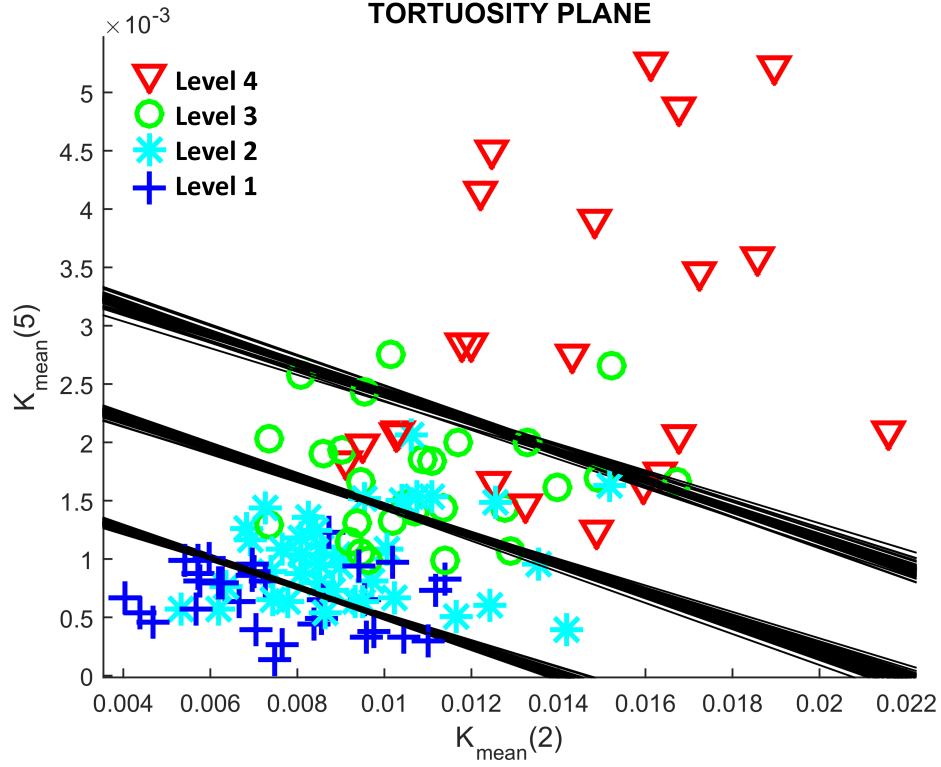


Figure 5.6 All images in IVCMI40 are projected onto the tortuosity plane after a leave-one-out cross-validation on unseen images. Markers indicate the majority voting ground truth tortuosity level. The estimated tortuosity level corresponds to the region within which an IVCMI image is mapped. During each cross-validation the best decision boundaries (shown in black) are computed based on the training set.

level, showing very low MSE errors. Moreover, different regions show different spreading on the TP, especially in region 2 (very low) and region 4 (very high), suggesting that intra-level tortuosity characteristics should be better investigated in future studies.

5.4 Conclusions

In this chapter I have introduced and discussed the tortuosity estimation module of the proposed fully automated framework for image-level tortuosity estimation. This module is based on a novel tortuosity estimation paradigm, capable of identifying the most discriminative tortuosity features and their combination, among the pool of multi-scale tortuosity measures (i.e. it does not pre-define which features determine tortuosity estimation).

Experiments carried out on 140 images from healthy and unhealthy subjects with different pathologies (IVCM140), show that our segmentation module for tortuous structures outperforms conventional methods based on the “locally-straight” assumption and learned appearance filters. The improvement in terms of segmentation is immediately transferred to tortuosity estimation performance. The tortuosity estimation module, based on multi-scale spline-based curvature estimation, performs considerably better than state-of-the-art *single-index* algorithms and addresses the main drawback of the previously proposed multi-scale-multi-window approach, i.e. the speed. In fact, the proposed solution is orders of magnitude faster than the previous one, and achieves better performance. This speed gain, combined with the hybrid segmentation solution, makes the system fast: 30s are required to analyse an IVCM image using MATLAB (R2014a) code on a machine equipped with Intel i7-4770 CPU at 3.4 GHz.

Visualising the predictions made by the proposed tortuosity estimation method on a two-dimensional plane, the tortuosity plane, offers various advantages compared with traditional approaches providing a single number (i.e., tortuosity level or index). First, it allows a finer (i.e., continuous) tortuosity scale compared to methods estimating tortuosity using a few levels. Compared to approaches based on indices (i.e., providing a continuous number), the TP provides a two-dimensional continuous tortuosity scale, thus allowing a better interpretation of the estimated tortuosity. Moreover, the contribution of each tortuosity measure into the final estimate becomes trivial (i.e., it is readily available on the axes of the plane). Overall, the TP allows a better tortuosity stratification and the possibility to identify sub-categories clustering in specific regions of the plane. Second, it visualises the level of confidence (i.e., reliability) of the estimated tortuosity simply and intuitively, using colour and also by the distance of a point from its closest boundaries, an aspect ignored so far. Importantly, the proposed method takes into account that the tortuosity grade of some images can be estimated with less confidence compared to others.

The comparison with three experienced observers who annotated the images independently shows, remarkably, that the proposed system matches or even exceeds their performance. This opens the possibility of analysing the large volumes of images needed for screen-

ing programs very efficiently, subject of course to further validation with much larger data sets.

In the next chapter, I will discuss ways to improve curvilinear structure modelling and segmentation, which could make the proposed automated tortuosity quantification framework more effective.

Chapter 6

Improving Curvilinear Structure Modelling and Segmentation

6.1 Introduction

In the previous chapters I have mainly dealt with corneal nerve fibres and neurites. Experimental results show that the proposed segmentation module achieves a level of detection performance such that tortuosity estimation matches or exceeds the level of performance of cornea specialists. With the aim of making the proposed system versatile, I have carried out experiments with other curvilinear structures such as blood vessels acquired with fundus camera and found that SCIRD tends to perform poorly on very thin structures (1 or 2 pixels wide)¹. In this chapter I discuss the reason why SCIRD tends to degrade its detection performance when dealing with very thin structures (i.e. retinal blood vessels captured at low resolution) and propose a new formulation of SCIRD, SCIRD-TS, which addresses this problem and improves the detection of very thin structures considerably.

The proposed hybrid segmentation approach makes use of multi-range context filters. Those filters are learned with K-means clustering due to the prohibitive cost of learning several hundreds of filters with strategies such as convolutional sparse coding, potentially

¹Curvilinear structures in IVC100, IVC140, BF2D and VC6 are typically wider, so SCIRD performs well.

leading to more discriminative filter banks. In this chapter I also discuss a novel acceleration approach to speed-up convolutional sparse coding filter learning, which reduces considerably the time needed to learn such filter banks. The benefits of reducing the training time go beyond improving the proposed segmentation module and could be adopted by state-of-the-art deep learning architectures (or DLAs) recently proposed.

6.2 Improving SCIRD

6.2.1 SCIRD for thin structures (SCIRD-TS)

Curvilinear structures such as blood vessels and neurites share appearance characteristics which can be easily modelled, rather than learned. In recent years, important efforts have been made in this regard and several HCFs have been proposed (e.g., Frangi [49], Gabor [123], OOF [83]). These methods assume that a curvilinear structure is “locally straight” and well contrasted. However, these assumptions are violated by structures such as blood vessels and neurites, appearing fragmented, showing some level of tortuosity or captured with low signal-to-noise ratio. As a consequence, detection performance may degrade significantly. I addressed these modelling issues in Chapter 4 by proposing a novel ridge detector, SCIRD, which adds curvature and contrast invariance to that of previous HCFs (i.e., scale, rotation and elongation).

In Chapter 4 (SCIRD), I model a curvilinear structure with a *curved-support* Gaussian function. Then, the curved ridge detection is obtained by measuring the second directional derivative along the gradient of each curved-support Gaussian. This derivation results in a ridge detector which consists of a ratio of first and second derivatives of the curved-support Gaussian function, thus leading to “0/0” indeterminate form in particular cases, e.g. when the first derivatives vanish. Unfortunately, this compromises the detection of thin structures, as shown qualitatively in Figure 6.1 and 6.2.

To address this limitation and avoid indeterminate forms, here I modify the derivation of the curved-support ridge detector. Specifically, instead of curving the curvilinear structure model (as done for SCIRD), I first derive a *straight* ridge detector. Then, I apply a non-

linear transformation to curve the ridge detector. This new ridge detector is therefore curved as SCIRD is, but when adopting straight filters (i.e. curvature is 0), it does not lead to indeterminate pixel values, as shown in Figure 6.1 (third row). This improves the detection of thin structures, as shown qualitatively in Figure 6.2.

Let us model a *straight* ridge-like structure by means of a multivariate zero-mean (n -D) Gaussian function with diagonal covariance matrix,

$$G(\boldsymbol{\varphi}; \boldsymbol{\sigma}) = \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^n \sigma_i^2}} \exp \left(- \sum_{i=1}^n \frac{\varphi_i^2}{2\sigma_i^2} \right) \quad (6.1)$$

where $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \dots, \varphi_n)$ represents a point in the $\{\boldsymbol{\varphi}\}$ coordinate system, and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \dots, \sigma_n)$ describes the standard deviation in each direction. A ridge *detector* can be obtained by measuring the contrast between the part inside and outside the ridge [49]. This can be achieved by measuring the second derivative with respect to the variables along which we observe the ridge-like profile. Using the separability property of the n -D Gaussian, one can compute the second derivative with respect to each variable and then combine the results (e.g. by summing up all the contributions). The second derivative of $G(\boldsymbol{\varphi}; \boldsymbol{\sigma})$ with respect to the variable φ_j has the form

$$G_{\varphi_j \varphi_j}(\boldsymbol{\varphi}; \boldsymbol{\sigma}) = G(\boldsymbol{\varphi}; \boldsymbol{\sigma}) \left[\frac{1}{\sigma_j^2} \left(\frac{\varphi_j^2}{\sigma_j^2} - 1 \right) \right]. \quad (6.2)$$

If we assume (without loss of generality) that the structure shows a ridge-like profile only with respect to the coordinate φ_j , the function $G_{\varphi_j \varphi_j}(\boldsymbol{\varphi}; \boldsymbol{\sigma})$ represents a ridge detector for *straight* structures. To extend this ridge detector to curved objects, we can consider a non-linear transformation $\mathcal{T} : \mathbb{R}^n \mapsto \mathbb{R}^n$ with $\mathcal{T}(\mathbf{x}) = \boldsymbol{\varphi} = (\varphi_1, \varphi_2, \dots, \varphi_n)$ of the form

$$\varphi_j = x_j + \sum_{i=1}^{n-1} k_{ji} x_i^2, \quad 2 \leq j \leq n \quad (6.3)$$

and $\varphi_1 = x_1$, where $k_{ji} \in \mathbb{R}$ and x_i are the coordinates of a point in the new $\{x\}$ coordinate system. In the 2-D case (i.e. $n = 2$), applying the transformation \mathcal{T} in Eq. (6.3) to

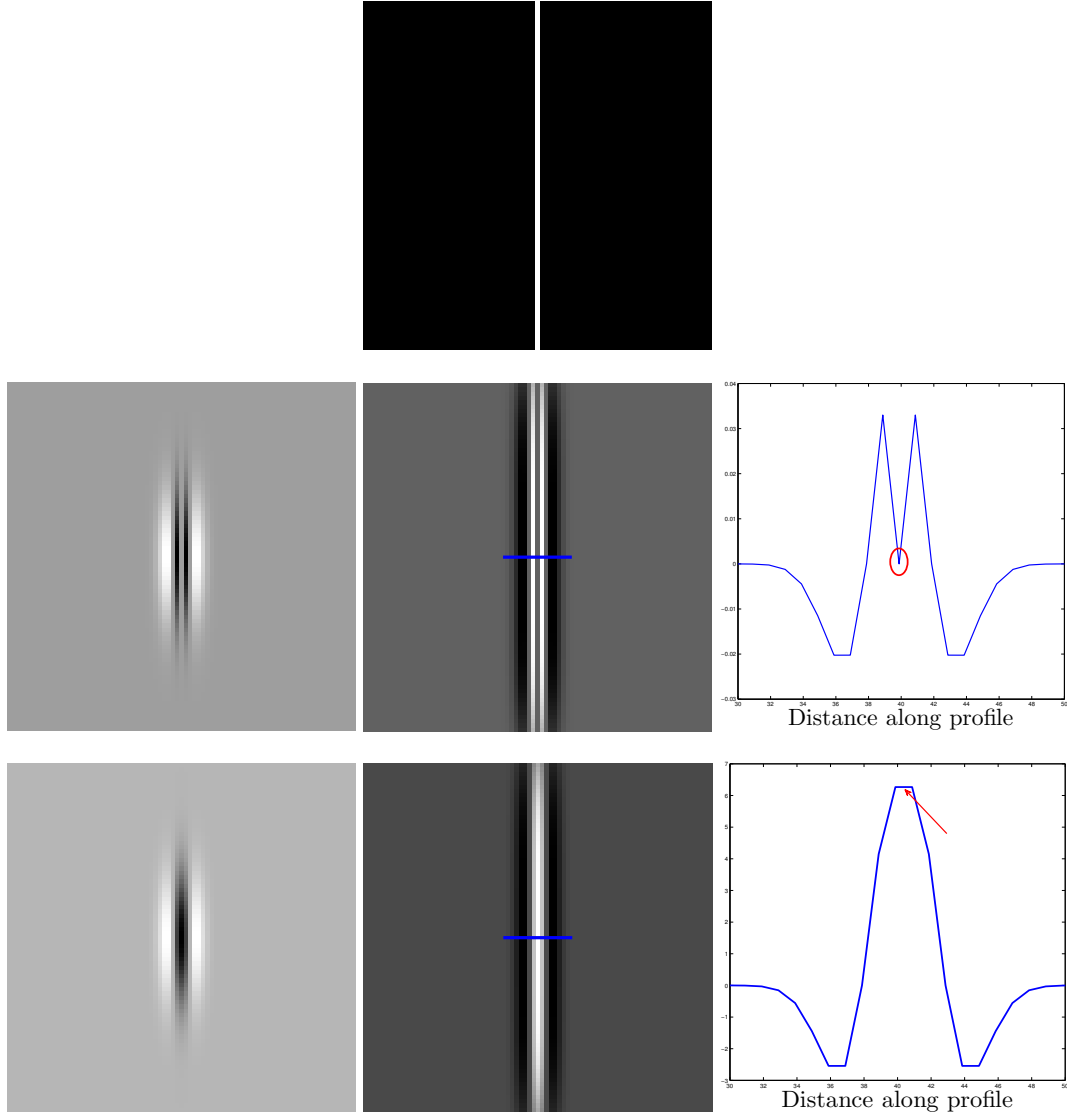


Figure 6.1 First row: ideal thin structure (1 pixel wide); second row, from left to right: SCIRD filter, SCIRD response and its cross-sectional profile along the blue line; third row, from left to right: SCIRD-TS filter, SCIRD-TS response and its cross-sectional profile along the blue line. Notice that while the SCIRD response is approximately 0 on the thin structure (i.e. SCIRD does not detect it), the SCIRD-TS one is maximum, hence leading to a correct detection.

$G_{\varphi_j \varphi_j}(\boldsymbol{\varphi}; \boldsymbol{\sigma})$ in Eq. (6.2), leads to the general form of a SCIRD-TS filter:

$$F(\mathbf{x}; \boldsymbol{\sigma}, k) = \frac{1}{\sigma_2^2 Z(\boldsymbol{\sigma})} \left[\frac{(x_2 + kx_1^2)^2}{\sigma_2^2} - 1 \right] \exp\left(-\frac{x_1^2}{2\sigma_1^2}\right) \exp\left(-\frac{(x_2 + kx_1^2)^2}{2\sigma_2^2}\right), \quad (6.4)$$

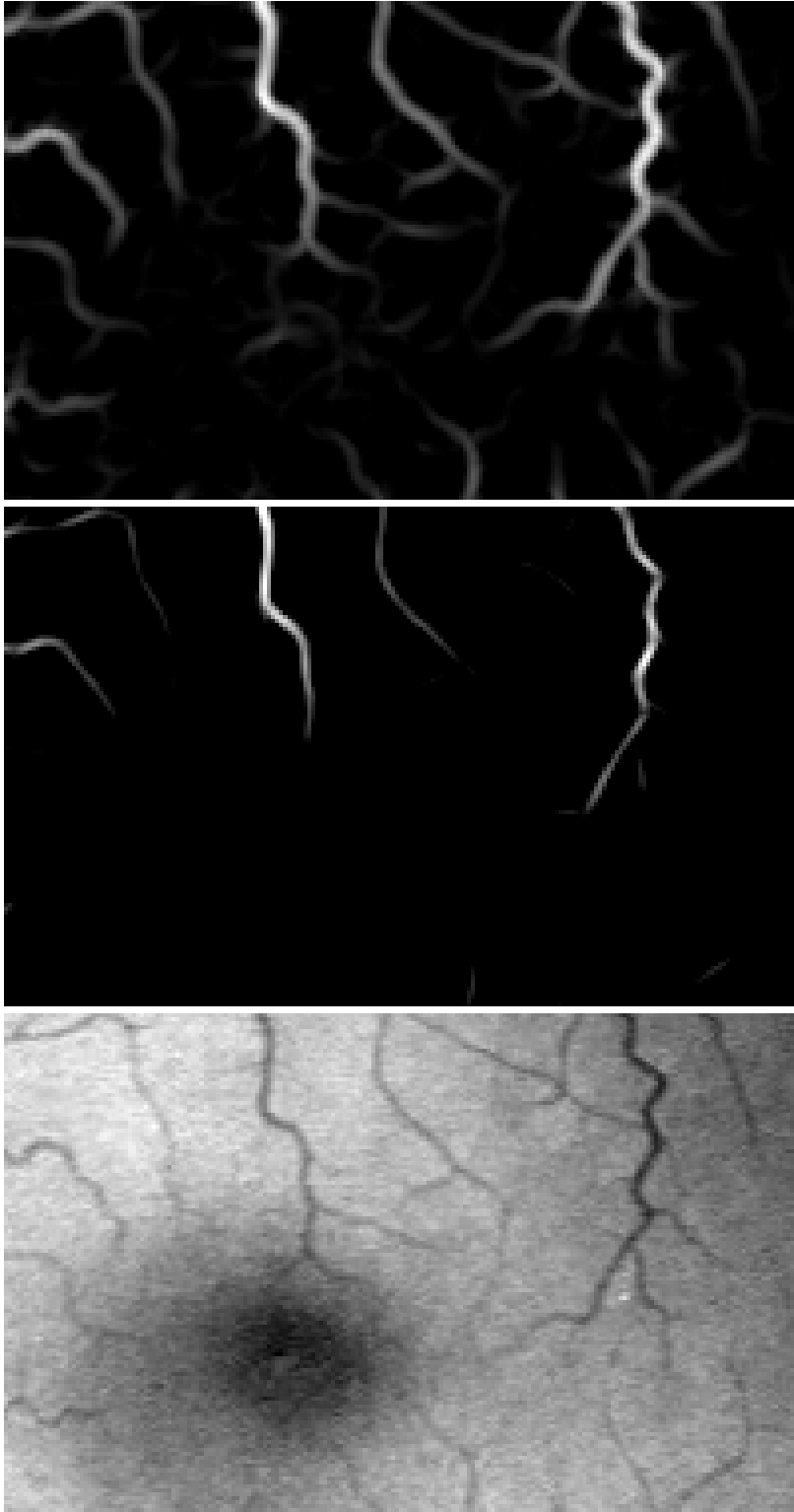


Figure 6.2 Detecting thin vessels. Left: original image patch (green channel) showing thin retinal blood vessels around the fovea; middle: enhancement using SCIRD; right: enhancement using the proposed SCIRD-TS. The thin vessels not enhanced by SCIRD are correctly enhanced by SCIRD-TS.

where k_{21} (curvature parameter) is indicated as k for compactness.

To make the ridge detector rotation invariant, SCIRD-TS filters can be simply rotated by θ , applying the rotation matrix to (x_1, x_2) . Therefore, we will indicate a SCIRD-TS filter as $F(\mathbf{x}; \boldsymbol{\sigma}, k, \theta)$.

A pre-defined convolutional filter bank can be generated by spanning the range of the free parameters σ_1 , σ_2 , k and θ . Similarly to SCIRD (Section 4.2.2), contrast normalisation can be adopted for SCIRD-TS as well. Then, the unsupervised version of SCIRD-TS can be simply obtained by computing the maximum projection over all the filter responses, while the supervised one can be built by using SCIRD-TS filter bank responses as input to a classifier (e.g. Random Forest).

6.2.2 Experiments and results

Data sets. I employed four benchmark data sets to validate SCIRD-TS detection performance. They include two of the most used data sets to validate retinal blood vessel segmentation, DRIVE (Section 3.2.5) and STARE (Section 3.2.6), and two data sets showing neurites, BF2D (Section 3.2.3) and VC6 (Section 3.2.4), used as benchmark in recent work [111, 121, 122].

Performance evaluation protocol. I adhere to the evaluation protocol adopted in [111, 121, 122], among others. Specifically, given the noticeable imbalance between true negatives (TNs) and the other measures of the contingency matrix, i.e. true positives (TPs), false negatives (FNs) and false positives (FPs)², I adopt PRCs to assess segmentation performance. I compare SCIRD-TS detection performance with widely used HCFs (i.e. Gabor [123], Frangi [49], OOF [83]) and SCIRD (Section 4.2.2).

Parameter setting. Parameters for SCIRD-TS, SCIRD and baseline methods were tuned independently to achieve their best performance on each data set, to provide a fair comparison. All the experiments were carried out on a laptop equipped with Intel i7-4702 CPU at 2.2GHz and 16GB RAM (MATLAB implementations, R2014a).

²The number of true background pixels is much higher than that of true vessel or neurite pixels in the images.

Results and discussion. Figure 6.3 shows the segmentation performance on the four data sets in terms of precision-recall curves for SCIRD-TS, SCIRD and state-of-the-art and widely used HCFs. The proposed SCIRD-TS outperforms SCIRD (and the other HCFs baselines) on the four data sets, as it detects thinner structures not detected by SCIRD. Notice that the biggest improvement is observed on DRIVE and STARE data sets (retinal blood vessels) which include a large number of very thin vessels. The improvement is less noticeable on the other two data sets, as the resolution with which curvilinear structures are imaged is higher, therefore there are not so many very thin vessels. Nevertheless, SCIRD-TS matches or exceeds the detection performance of SCIRD on those data sets, thus suggesting that better detection performance on very thin structures does not worsen the detection of large ones.

6.3 Improving unsupervised filter learning

Most of the existing methods for automated curvilinear structure segmentation rely on HCFs designed to model local geometric properties of ideal *tubular* shapes. Today research is moving towards DLAs given their excellent results on several challenging tasks, as shown by Bengio et al. [17], Kavukcuoglu et al. [75], Kontschieder et al. [76], and Krizhevsky et al. [78], among others. In medical image analysis, DLAs have been used for segmentation by Brebisson and Montana [21], Ciresan et al. [34, 35], Li et al. [90] and Kamnitsas et al. [73], among others, but also for other applications, such as predicting Alzheimer’s disease by Payan and Montana [107] (refer to Schmidhuber [117] for a comprehensive review).

Experimental results show that lower layers of DLAs with convolutional structure (e.g. convolutional neural networks, or CNN) tend to learn a subset of filters similar to well-known HCFs (e.g. Gabor filters, see [24, 63, 75, 78, 89, 121]). This is also the case for convolutional sparse coding (henceforth, CSC) as shown in Figure 6.4. Therefore, I argue that employing such complex architectures to learn filters similar to HCFs is inefficient; a more efficient approach would be learning *only* appearance characteristics not included in the hand-crafted models. On the other hand, HCFs often require manual parameter tuning

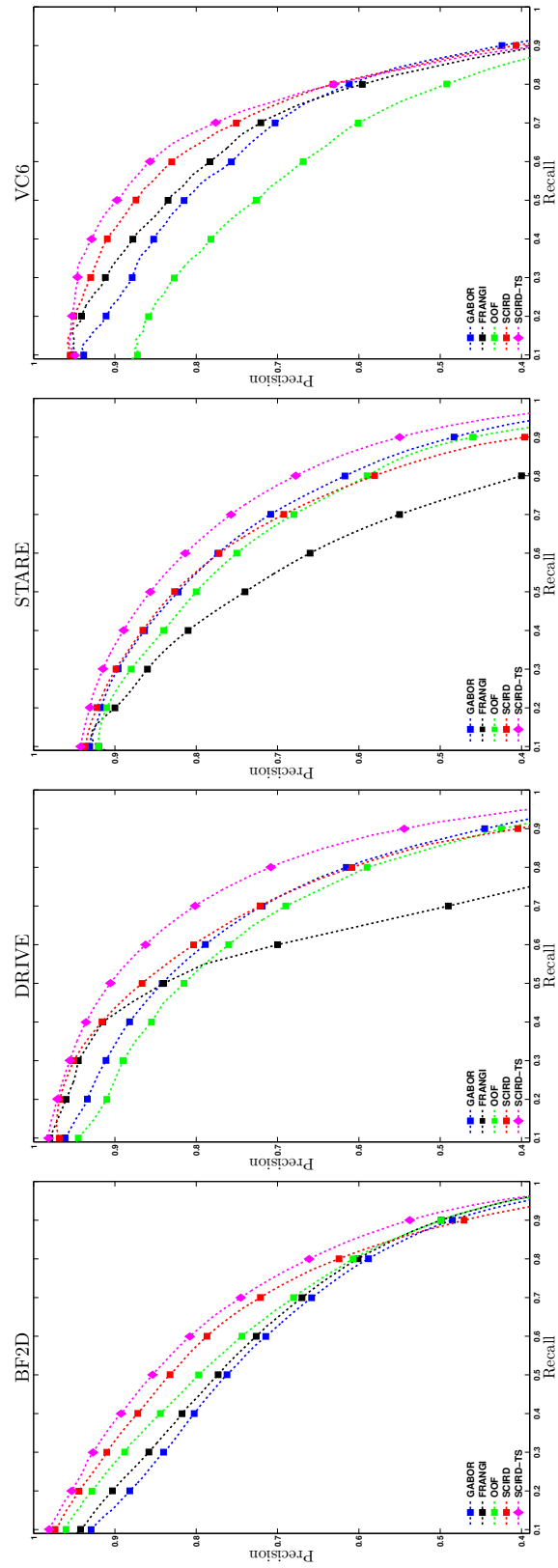


Figure 6.3 Performance evaluation in terms of precision-recall curves (pixel-level segmentation) for several HCFs on BF2D, DRIVE, STARE and VC6 data sets.

which does not guarantee optimal performance, while DLAs are capable of finding the best setting automatically.

Recently, an auto-context framework (multi-layer) based on unsupervised filter learning has been shown to outperform CNN and modifications [51] on curvilinear structure segmentation in the medical domain [120, 122]. The framework proposed in [120, 122] relies on filters learned through CSC [111, 121], but learning them is very time-consuming as reported in [111] (several days to learn 121 filters using MATLAB code and state-of-the-art machines). Therefore, the filter bank learned at the first layer is kept unchanged across the other ones, due to the prohibitive cost of learning layer-specific filter banks [122]. This limitation is particularly relevant for medical imaging applications, where the visual appearance of curvilinear structures may vary significantly and the range of acquisition modalities may lead to different image characteristics in terms of contrast and noise. As a consequence, re-training could be necessary to achieve good performance.

Motivated by the above and inspired by the observation that filters learned by CSC for curvilinear structure segmentation are often similar to well-known HCFs, I propose an efficient approach to learning CSC filters.

This work differs fundamentally from recent acceleration methods like those reported by Heide et al. [63], Bristow et al. [24], and Bao et al. [12, 13], which rely on efficient mathematical formulations to solve the CSC *optimisation* problem. Such methods typically initialise filters with random values or by a discrete cosine transform (henceforth, DCT). While this solution is general, it does not exploit prior knowledge about the target curvilinear structure and its appearance.

Unlike previous methods, I achieve CSC acceleration by a novel *warm-start initialisation* strategy based on SCIRD-TS. Specifically, the proposed warm-start strategy identifies the optimal set of initial filters from a large amount of HCFs generated by spanning the range of parameters related to the structures of interest. It is worth noting that setting the ranges for HCF parameters is very intuitive, as they represent geometric properties of the target structure and their effects can be checked visually. These filters are then *refined* by using CSC to incorporate specific properties of the structures (e.g. retinal blood vessels,

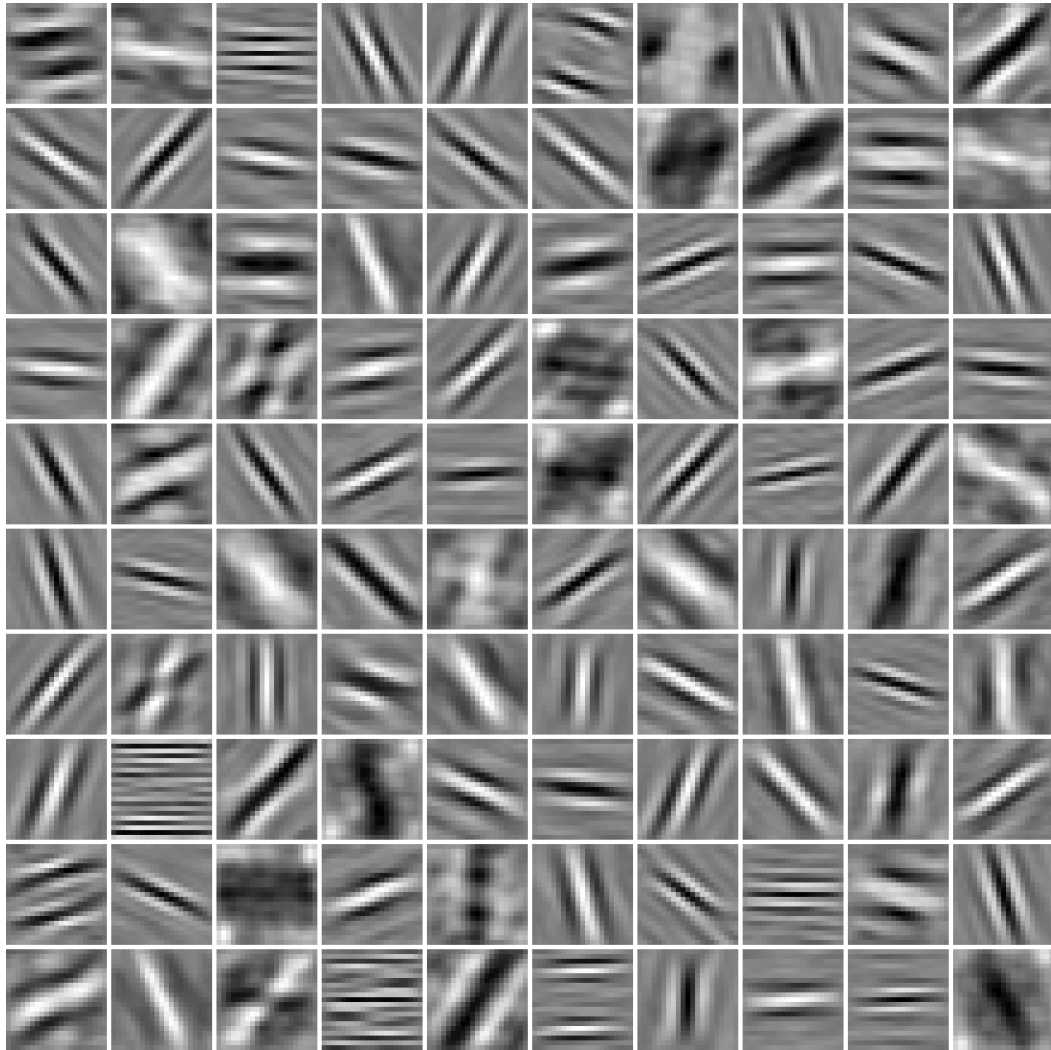


Figure 6.4 A filter bank learned using convolutional sparse coding with random initialisation. The DRIVE data set was used for this experiment.

neurites) of a specific data set. Intuitively, the speed-up is achieved by learning only the “properties” which have not been modelled and by refining the ones already modelled (e.g. width or elongation).

Importantly, any previously proposed CSC solver for filter learning could be adopted in our framework (e.g. [14, 24, 63]). For this reason, the proposed acceleration method could be *combined* with state-of-the-art (and future) fast CSC solvers.

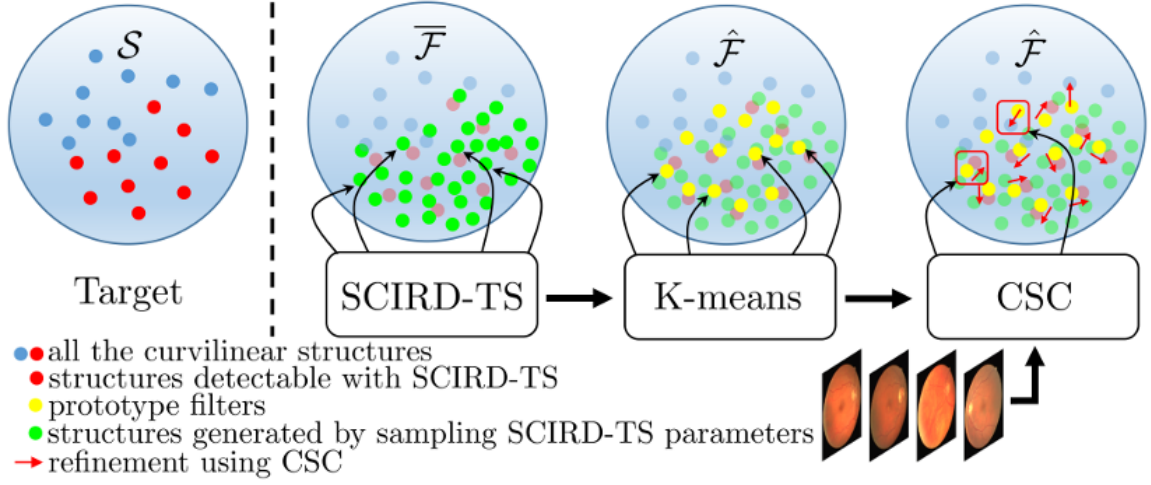


Figure 6.5 Block diagram of the proposed method. Notice that all the curvilinear structures, represented in the space \mathcal{S} as blue and red dots, are copied in the other spaces as well (to show their original position), and they are shown in light blue and light red, respectively.

6.3.1 Optimal warm-start strategy

Let $\mathcal{S} \subseteq \mathbb{R}^p$ be the space of all the curvilinear structures in a particular data set, and assume that a subset s of them can be detected by using SCIRD-TS filters in the space $\mathcal{F} \subseteq \mathbb{R}^p$ (Figure 6.5). The parameter ranges of these SCIRD-TS filters can be estimated easily, e.g. by visual inspection of the curvilinear structures in \mathcal{S} . Sampling such parameter ranges with regular spacing (i.e., fixed sampling step) guarantees better approximations of the filters in \mathcal{F} as the sampling step δ vanishes. Let $\overline{\mathcal{F}} \in \mathbb{R}^p$ be the space generated by such sampling procedure. So, the first step of the proposed warm-start strategy consists of generating t SCIRD-TS filters in $\overline{\mathcal{F}}$ ($t > s$, in general) by sampling regularly and densely (i.e., small sampling step) its parameter ranges.

Using the entire set of SCIRD-TS filters generated in the previous step is clearly infeasible ($t > 20,000$ with our parameter setting). So, we need to reduce the cardinality of $\overline{\mathcal{F}}$ and represent it with filters in a new space $\hat{\mathcal{F}}$ with a much lower cardinality $K \ll t$, while still preserving a good representation of $\overline{\mathcal{F}}$ (hence of \mathcal{F}). A key requirement for the success of sparse coding dictionary learning strategies is building *incoherent* dictionaries (e.g. [13]). The mutual *incoherence* of a dictionary D can be defined as

$$\mu(D) = \min_{i \neq j} \|\mathbf{d}_i - \mathbf{d}_j\|_2^2, \quad (6.5)$$

where \mathbf{d}_i and \mathbf{d}_j are two different dictionary elements (or atoms) arranged as column vectors. So, a high value of $\mu(D)$ for the learned dictionary is desirable. Moreover, since our overall target is to accelerate CSC, the cardinality reduction should be fast, so that most of the training time is spent on the CSC phase. Of course, sampling uniformly and sparsely SCIRD-TS parameter ranges would be fast, but it would not guarantee high dictionary incoherence.

The compression approach I adopt here to identify the set of K *prototype filters* which represent optimally (in the sense of minimising the quantisation error) the original SCIRD-TS space $\overline{\mathcal{F}}$ is K -means clustering using Euclidean distance³. K -means clustering offers: (1) an *optimal* compression approach for any chosen K , thus meeting the requirement of good representation of the original SCIRD-TS space; (2) the desirable high mutual incoherence (i.e. high inter-cluster Euclidean distance)⁴; (3) a fast compression algorithm (run time negligible compared to the CSC phase). So, if we indicate with $\mathbf{f}^{(i)}$ the i -th SCIRD-TS filter in $\overline{\mathcal{F}}$ ($\mathbf{f}^{(i)}$ is $F(\mathbf{x}; \boldsymbol{\sigma}, k, \theta)$ in Eq. (6.4) arranged as a column vector), the second step of the proposed warm-start strategy consists of solving the optimization problem

$$\operatorname{argmin}_{D, \mathbf{c}} \sum_i \left\| D\mathbf{c}^{(i)} - \mathbf{f}^{(i)} \right\|_2^2 \quad (6.6)$$

subject to $\|\mathbf{c}^{(i)}\|_0 \leq 1, \forall i = 1, \dots, m_D$ and $\|\mathbf{d}^{(j)}\|_2 = 1, \forall j = 1, \dots, K$, where $\mathbf{c}^{(i)}$ is the code vector related to the i -th original SCIRD-TS filter $\mathbf{f}^{(i)}$, and $\mathbf{d}^{(j)}$ is the j -th column of the dictionary D of prototype filters (examples in Figure 6.7, first column). In our experiments, we adopt the fast K-means optimisation algorithm proposed by Coates and Ng in [36]⁵. Careful seeding discussed in [8] is used to initialise the clusters.

³I adopt the same distance used for the CSC phase.

⁴It is worth noting that optimising Equation (6.6) is not exactly the same as optimising Equation (6.5). In fact, Equation (6.6) guarantees (locally) optimal distance among *all* cluster centres/filters (i.e. on average), while Equation (6.5) requires that the minimum distance among the filters is high. Nevertheless, using K-means often leads to filter banks with high incoherence.

⁵Notice that this algorithm does not guarantee convergence to the global minimum but to a local one, so the compression is *locally* optimal.

6.3.2 Refining the prototype filters by CSC

I refine the filter bank obtained with the warm-start strategy by CSC. Specifically, I optimise the following objective function [111]:

$$\operatorname{argmin}_{\substack{\{D^{(j)}\} \\ \{M_i^{(j)}\}}} \sum_{i=1}^N \left(\left\| P_i - \sum_{j=1}^K D^{(j)} * M_i^{(j)} \right\|_2^2 + \lambda \sum_{j=1}^K \|M_i^{(j)}\|_1 \right), \quad (6.7)$$

where P_i is the i -th original image patch to reconstruct (N patches in total), $D^{(j)}$ is the j -th refined filter (K filters in total), $M_i^{(j)}$ can be regarded as the j -th component (map) of the representation related to P_i and λ is the sparsity (regularization) parameter. Filters, original image patches and representation maps are arranged as matrices. The symbol $*$ indicates convolution.

In essence, the goal of this CSC optimisation is to minimise the total reconstruction error computed by approximating each original image patch using the current filter bank. The reconstruction is obtained by finding a sparse representation of the current patch (the second term in Eq. (6.7) penalises the ℓ_1 -norm of each component of the representation). Since the objective in Eq. (6.7) is not convex, several optimisation strategies can be employed. For instance, Rigamonti et al. [111] adopted a proximal algorithm, *ISTA* (Iterative Shrinkage Thresholding Algorithm) [11, 106]. To speed-up the optimisation, I adopt a faster proximal method, *FISTA* [14]. Moreover, I compute the high number of convolutions in the Fourier domain by exploiting fast Fourier transform algorithms. Finally, I adopt a batch-based optimisation strategy as done, for instance, in [87, 128].

6.3.3 Impact of the warm-start strategy on CSC optimisation

I provide a brief analysis of the computational complexity of CSC optimisation, in terms of number of multiplications, to better investigate the impact of the proposed warm-start strategy on the running time.

Let $I_1 \in \mathbb{R}^{r_1 \times c_1}$ and $I_2 \in \mathbb{R}^{r_2 \times c_2}$ be two images (or patches) we want to convolve. Due to the high number of convolutions involved in the CSC optimisation, I compute them in the Fourier domain, hence requiring the following steps:

1. Padding I_1 and I_2 with zeros so that they have the same size $r_3 \times c_3$, where r_3 and c_3 are the closest powers of 2 larger than $r_1 + r_2 - 1$ and $c_1 + c_2 - 1$, respectively;
2. Computing the Fourier transform (DFT) of the two images;
3. Multiplying the DFTs of the two images;
4. Computing the inverse Fourier transform (IDFT) of the result.

Considering that a DFT (and also an IDFT) requires $6r_3c_3 \log_2(r_3c_3)$ real multiplications [121], and that a complex multiplication requires 3 real multiplications, a single convolution would require $3r_3c_3(6\log_2(r_3c_3) + 1)$ multiplications.

The fast proximal method (FISTA) I adopt to optimise Eq. (6.7) alternates between the optimisation w.r.t. the K filters ($D^{(j)}$) and the maps ($M_i^{(j)}$) for each patch P_i :

Optimisation w.r.t. the filters. This can be obtained by gradient descent, which amounts to computing K convolutions between the residual error of reconstruction and the related K maps, as the second term of Eq. (6.7) vanishes [32]. The total number of multiplications needed to perform this step is therefore⁶ $3Kr_3c_3(6\log_2(r_3c_3) + 1)$.

Optimisation w.r.t. the maps. From a computational complexity perspective, this step requires the computation of the gradient of the first term in Eq. (6.7) w.r.t. the maps $M_i^{(j)}$ and a soft-thresholding (proximal operator of the l_1 norm [32, 106]). Again, the gradient can be computed efficiently by convolving the K filters with the residual error of reconstruction [32], hence requiring $3Kr_3c_3(6\log_2(r_3c_3) + 1)$ multiplications. In addition, the K soft-thresholding operations require Kr_3c_3 multiplications [24].

Since we optimise over N patches (also called “mini-batch” in batch-based optimisation strategies [87, 102, 128]) and iterate several times (every pass over all the N patches is denoted as “epoch”, N_e), the total number of multiplications required to optimise Eq. (6.7)

⁶One could pre-compute the DFT of the residual error and reduce the number of multiplications further.

is:

$$N_e \times N \times [6Kr_3c_3(6\log_2(r_3c_3) + 1) + Kr_3c_3]. \quad (6.8)$$

The number of patches N , the number of filters K and the dimension of the filters are application-dependent. Once the optimisation algorithm is fixed (FISTA, in this case), the only other parameter which could have a significant impact on the complexity is the number of epochs N_e (multiplicative factor). I demonstrate experimentally in the next section that initialising CSC with the proposed warm-start strategy reduces N_e (and often achieves lower reconstruction errors, thus potentially leading to more discriminative filter banks).

6.3.4 Experiments and results

Data sets. I employed four benchmark data sets to validate the proposed CSC acceleration strategy. They include two of the most popular data sets to validate retinal blood vessel segmentation, DRIVE [125] and STARE [66], and two data sets showing neurites, BF2D and VC6, used as benchmark in recent work [111, 121, 122]. Poor and variable contrast, low-resolution, non-uniform illumination, structure fragmentation, irregularities in the staining process (VC6), confounding non-target structures (e.g. optic disk, exudates and haemorrhages in DRIVE; blob-like structures in BF2D and VC6) make these data sets particularly challenging for automatic segmentation.

Performance evaluation. Since the CSC optimisation problem aims to find a sparse representation for each original image patch minimising the total reconstruction error, I first assessed the performance in terms of reconstruction error and time to convergence. Then, I evaluated segmentation performance.

- *Reconstruction error and time to convergence.* For these experiments, I randomly sampled 1,000 49×49 image patches from the original images (i.e. the “batch”) from the training set of DRIVE, BF2D and VC6 separately and measured the total reconstruction error against the number of epochs⁷. For STARE, I excluded the 20 manually segmented images used for assessing segmentation performance and car-

⁷In batch-based optimisation strategies, an epoch represents one pass over the entire batch.

ried out this experiment on the 377 images left. I compared the performance of the proposed initialisation strategy against the random one, adopted in most of the related work, e.g. [24, 63, 111, 121], and DCT-based one, adopted in [12, 13]. I used the same batch for the proposed method and the baselines for fair comparison. To assess the influence of the dictionary size on the total reconstruction error I ran experiments for banks including 49, 100 and 144 learned filters.

- *Segmentation.* To assess segmentation performance, I convolve each image with the K learned filters and represent each pixel with the K local responses (i.e. K -D feature vector). Then, I give this feature vector as input to a random forest classifier to infer the probability of each pixel of belonging to a curvilinear structure. For DRIVE, BF2D and VC6, the training set was formed by pixel samples from the provided training images; for STARE, I adopted a leave-one-out cross-validation on the 20 images manually segmented, as typically done in the literature (e.g. [89, 123]). I adhere to the evaluation protocol adopted in [111, 121, 122], and compute PRCs and AUPRC to assess segmentation performance. In addition to the baselines adopted above (i.e. CSC with random and DCT initialisation), I compared the proposed method's performance with widely used HCFs (i.e. Gabor [123], Frangi [49], OOF [83]), SCIRD, and the combination method proposed by Rigamonti et al.[111].

Parameter setting. I report here the setting for the warm-start strategy, the CSC phase and the classifier.

- *Warm-start strategy.* Parameter ranges for generating the large SCIRD-TS filter bank were set manually by visually inspecting DRIVE training images, with the idea of covering a suitable range in terms of width, elongation, curvature and rotation resolution. I adopted a conservative setting (i.e., wide ranges and high resolution) without careful tuning or specific optimisation. In particular, $\sigma_1 = [1, 10]$ with step 0.5, $\sigma_2 = [1, 10]$ with step 0.5 (filters are forced to be elongated, i.e. filters with $\sigma_2 > \sigma_1$ are discarded), $k = [-0.1, 0.1]$ with step 0.025 and $\theta = [15, 180]$ with step 15 degrees. To test the generalisation of these settings, I adopted them for BF2D, VC6 and STARE as

well, although they contain different curvilinear structures (neurites vs retinal blood vessels) and resolution. I set the number of K -means iterations to 100, although a few tens are typically sufficient (with negligible impact on the total time to convergence of the proposed acceleration strategy). I assessed the influence of the number of filters (i.e., K) on the reconstruction performance using $K \in \{49, 100, 144\}$. For comparison, the maximum number of CSC filters learned in [120] (the current benchmark on DRIVE) is 121.

- *CSC phase.* When random initialisation is used, setting the sparsity parameter λ manually is not trivial. In fact, low values tend to produce noisy filters, whereas high ones lead to a slow convergence. I found $\lambda = 2$ to yield good results on the DRIVE data set; I investigated the impact of different λ values and report the results below. To test robustness, I used the same value for BF2D, VC6 and STARE as well.
- *Classifier.* I trained a RF using 144-D feature vectors (i.e. number of learned filters $K = 144$) with 100 trees for each data set, to achieve a good compromise between segmentation performance and processing time. Each tree's depth was set automatically, by evaluating the out-of-bag error during training. I randomly sampled 200,000 training instances from the training partition of each data set to build the related RF model.

I adopted the same filter size used in [111, 120, 121], i.e. 21×21 pixels, for all the data sets, as size was not found to affect performance significantly on the data sets used. All the experiments were carried out on a laptop equipped with Intel i7-4702 CPU at 2.2GHz and 16GB RAM (MATLAB R2014a).

Reconstruction error and time to convergence. In Table 6.1 I report the total time to convergence for CSC using the acceleration method and the baselines, for each data set and dictionary size. I observe that (1) the time to run our warm-start strategy is negligible compared to the total time to run CSC (i.e. a few seconds against tens of minutes); (2) the proposed CSC acceleration takes much less time to obtain discriminative filter banks than conventional CSC initialisation strategies, e.g. up to 82% less time, when 144 filters are

learned. Remarkably, the proposed acceleration strategy does not compromise performance either in terms of reconstruction error or segmentation performance. Figure 6.6 shows the total reconstruction error against the number of epochs needed to achieve convergence for the proposed initialisation method and the baselines, for the four data sets and different dictionary size. I notice that (1) the warm-start strategy based on SCIRD-TS achieves both the lowest total reconstruction error and the fastest convergence on each data set and for each dictionary size⁸; (2) initialising the filter bank with DCT (as done in [12, 13]) does not lead to either faster convergence or lower reconstruction error compared to random initialisation, for data sets including curvilinear structures⁹; (3) although the adopted SCIRD-TS parameters were set using DRIVE training images, the total reconstruction error on BF2D, VC6 and STARE is always lower, and sometimes substantially, than random initialisation, suggesting good generalisation. **Reconstruction error and time to convergence.** In Table 6.1 I report the total time to convergence for CSC using the acceleration method and the baselines, for each data set and dictionary size. I observe that (1) the time to run our warm-start strategy is negligible compared to the total time to run CSC (i.e. a few seconds against tens of minutes);

⁸In Figure 6.6 (“DRIVE - 144 FILTERS”) random initialisation achieves slightly less reconstruction error, with a substantially higher number of epochs.

⁹Bao et al. [12, 13] adopted DCT initialisation for general purpose applications, such as image compression.

Table 6.1 Total time to convergence (in minutes) for the CSC phase initialised with the proposed method (Prop.), and the baselines. In brackets, the proposed warm start processing time (in seconds).

DRIVE	Number of learned filters			BF2D	Number of learned filters		
Method	49	100	144	Method	49	100	144
Random	167	458	1085	Random	152	290	1062
DCT	167	242	2049	DCT	198	418	1474
Prop.	51(6)	106(11)	195(16)	Prop.	58(6)	141(11)	247(16)

VC6	Number of learned filters			STARE	Number of learned filters		
Method	49	100	144	Method	49	100	144
Random	132	374	467	Random	120	291	466
DCT	345	734	743	DCT	313	397	751
Prop.	54(6)	117(11)	203(16)	Prop.	61(6)	94(11)	159(16)

(2) the proposed CSC acceleration takes much less time to obtain discriminative filter banks than conventional CSC initialisation strategies, e.g. up to 82% less time, when 144 filters are learned. Remarkably, the proposed acceleration strategy does not compromise performance either in terms of reconstruction error or segmentation performance. Figure 6.6 shows the total reconstruction error against the number of epochs needed to achieve convergence for the proposed initialisation method and the baselines, for the four data sets and different dictionary size. I notice that (1) the warm-start strategy based on SCIRD-TS achieves both the lowest total reconstruction error and the fastest convergence on each data set and for each dictionary size¹⁰; (2) initialising the filter bank with DCT (as done in [12, 13]) does not lead to either faster convergence or lower reconstruction error compared to random initialisation, for data sets including curvilinear structures¹¹; (3) although the adopted SCIRD-TS parameters were set using DRIVE training images, the total reconstruction error on BF2D, VC6 and STARE is always lower, and sometimes substantially, than random initialisation, suggesting good generalisation.

Figure 6.7 illustrates how the initial filter banks generated using the proposed warm-start strategy were refined by the adopted CSC approach on each data set. A large subset of filters is left unchanged or refined lightly (in terms of width and elongation, for instance), while other filters are modified significantly to reduce the reconstruction error and compensate for the part HCFs are not capable to model. This observation confirms the underlying hypothesis that a well-designed HCF bank includes already a large portion of the filters suitable for curvilinear structures segmentation in the medical domain, and that the proposed approach (optimal warm-start) obtains highly discriminative filter banks in a more efficient way compared to conventional initialisation.

Segmentation. Figure 6.8 shows the segmentation performance on the four data sets in terms of PRCs for state-of-the-art and widely used HCFs (i.e. Gabor [123], Frangi [49], OOF [83]), SCIRD, SCIRD-TS, the combination approach proposed by Rigamonti et al.

¹⁰In Figure 6.6 (“DRIVE - 144 FILTERS”) random initialisation achieves slightly less reconstruction error, with a substantially higher number of epochs.

¹¹Bao et al. [12, 13] adopted DCT initialisation for general purpose applications, such as image compression.

Table 6.2 Comparison in terms of AUPRC, F-measure, Jaccard index and training time (in minutes), between random, DCT-based and the proposed initialisation strategy (denoted as “Ours”).

DRIVE		Performance measure			
Method		AUPRC	F-measure	Jaccard	Time
Random		0.85	0.77	0.62	1085
DCT		0.84	0.76	0.61	2049
Ours		0.87	0.79	0.64	195

BF2D		Performance measure			
Method		AUPRC	F-measure	Jaccard	Time
Random		0.83	0.77	0.62	1062
DCT		0.83	0.76	0.61	1474
Ours		0.84	0.76	0.62	247

VC6		Performance measure			
Method		AUPRC	F-measure	Jaccard	Time
Random		0.81	0.74	0.59	467
DCT		0.77	0.70	0.54	743
Ours		0.83	0.76	0.62	203

STARE		Performance measure			
Method		AUPRC	F-measure	Jaccard	Time
Random		0.84	0.75	0.58	466
DCT		0.83	0.74	0.57	751
Ours		0.86	0.77	0.60	159

[111] and CSC initialised with random initialisation (as done by [24, 63, 111, 121, 122]), DCT initialisation (as done by [12, 13]) and the proposed warm-start strategy. First, due to their modelling limitations and suboptimal parameter setting, HCFs are outperformed by methods based on discriminative filter learning. Second, precision-recall curves suggest that our acceleration strategy leads to filter banks matching or even exceeding the segmentation performance of CSC strategies initialised randomly or with general purpose HCFs (i.e. DCT), while converging in much less time. This is confirmed by quantitative results in terms of AUPRC, F-measure, Jaccard Index (aka Intersection Over Union, or IOU) and time needed to converge reported in Table 6.2. Qualitative comparisons (probability maps) with the best performing baseline (i.e. random initialisation) are reported in Figure 6.9.

Table 6.3 Influence of the sparsity parameter λ on the segmentation performance (AUPRC) of the proposed initialisation method (*Prop.*) and the best baseline method (*Random*) on DRIVE.

Init. method	λ		
	0.2	2	20
Random	0.8418	0.8515	0.8461
Prop.	0.8638	0.8676	0.8655

I investigated the influence of the sparsity parameter (λ) on the segmentation performance when the random and the proposed initialisation strategy are employed. Specifically, I repeated the experiments increasing and decreasing λ by a factor 10 compared to the adopted setting (i.e. $\lambda = 0.2$ and $\lambda = 20$, respectively) on the DRIVE data set. Experimental results (Table 6.3) suggest that CSC initialised with our warm-start strategy is more robust against this critical parameter setting, compared to random initialisation, an important advantage in terms of adaptation to different data sets.

It is worth noting that this segmentation pipeline is single-layer, yet it achieves the same level of performance as the multi-layer architecture proposed by Sironi et al. [122] on DRIVE (F-measure = 0.79); the latter is based on CSC filter banks leveraged by an auto-context regression pipeline recently improved by a post-processing strategy and shown to achieve state-of-the-art segmentation performance [120]. However, the authors report that learning a different convolutional filter bank for each layer of this auto-context architecture is “prohibitively expensive” [122], hence they learn a single filter bank (121 filters) and use it for all the layers. Given the speed-up obtained by using the proposed acceleration strategy (without performance degradation for reconstruction and segmentation), (1) a convolutional filter bank could be learned for each layer to model higher-order properties of curvilinear structures and potentially improve segmentation performance; (2) alternatively, the proposed acceleration strategy could significantly reduce its training time and therefore speed-up adaptation to other data sets.

6.4 Conclusions

In this chapter I have discussed methods to improve modelling and segmentation of curvilinear structures, with particular emphasis on the tortuous ones (target structures). Specifically, I have motivated the limit of the detection performance of SCIRD (indeterminate forms at the middle of straight filters) and proposed a novel formulation that rectifies this problem. Then, driven by experimental observations on the filter banks learned by state-of-the-art algorithms for supervised and unsupervised filter learning, I have proposed a novel approach to accelerate CSC for filter learning. The benefits of speeding up CSC could be directly employed to learn more discriminative multi-range context filters in the proposed segmentation module (although the current results seem to suggest that tortuosity estimation performance is already at the level of cornea specialists). More importantly, this acceleration could unlock the potential of DLAs based on auto-context, among the current state-of-the-art curvilinear structure segmentation approaches.

In the next chapter I will briefly summarise this thesis, discuss the limitations of the proposed tortuosity quantification framework and suggest potential solutions to explore in the future.

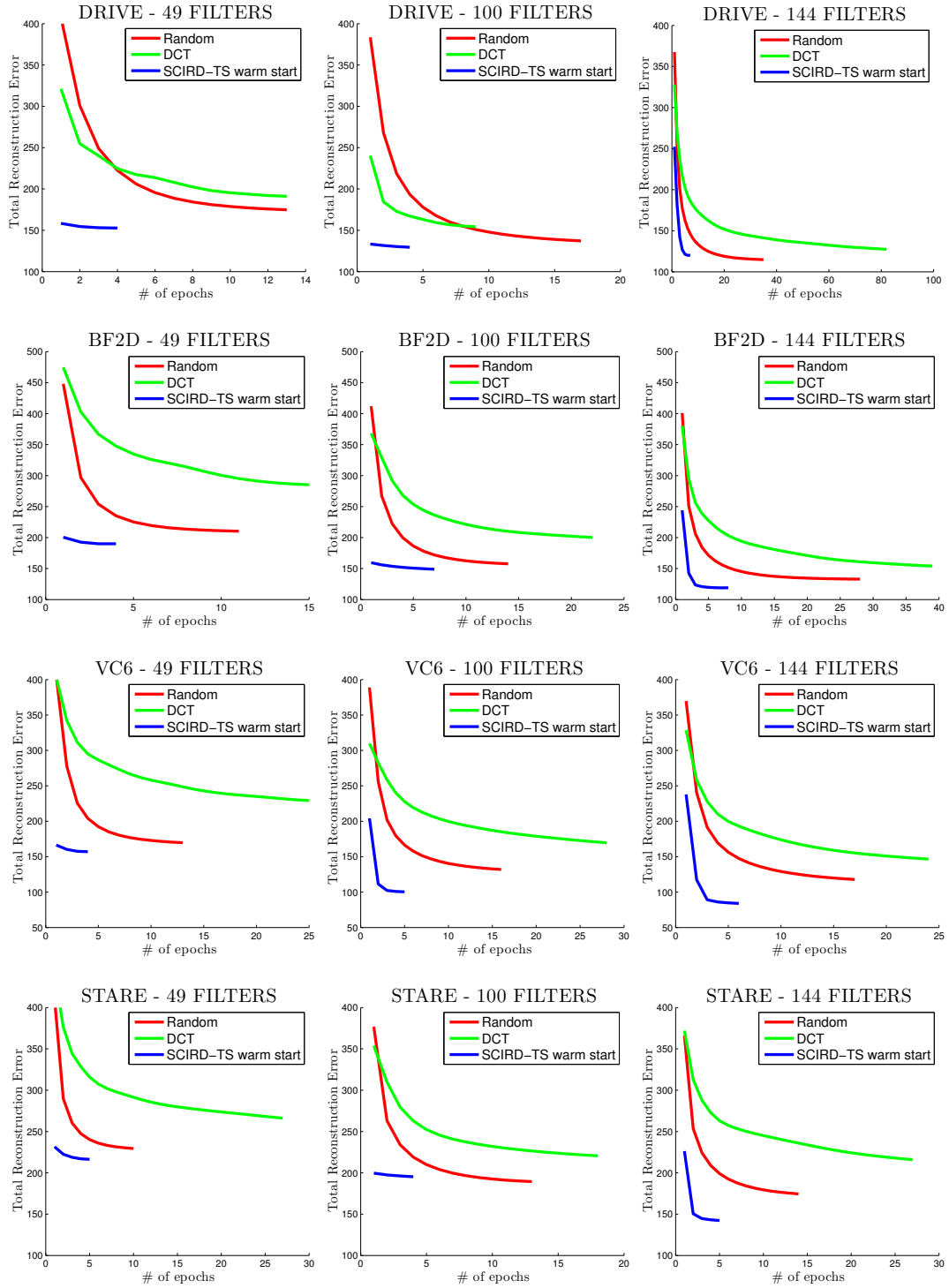


Figure 6.6 **Experiments: reconstruction error and time to convergence.** Performance evaluation in terms of total reconstruction error for CSC with random, DCT and SCIRD-TS initialisation. Each row shows the influence of the dictionary size on the total reconstruction error, for each data set. Optimisations were stopped at convergence.

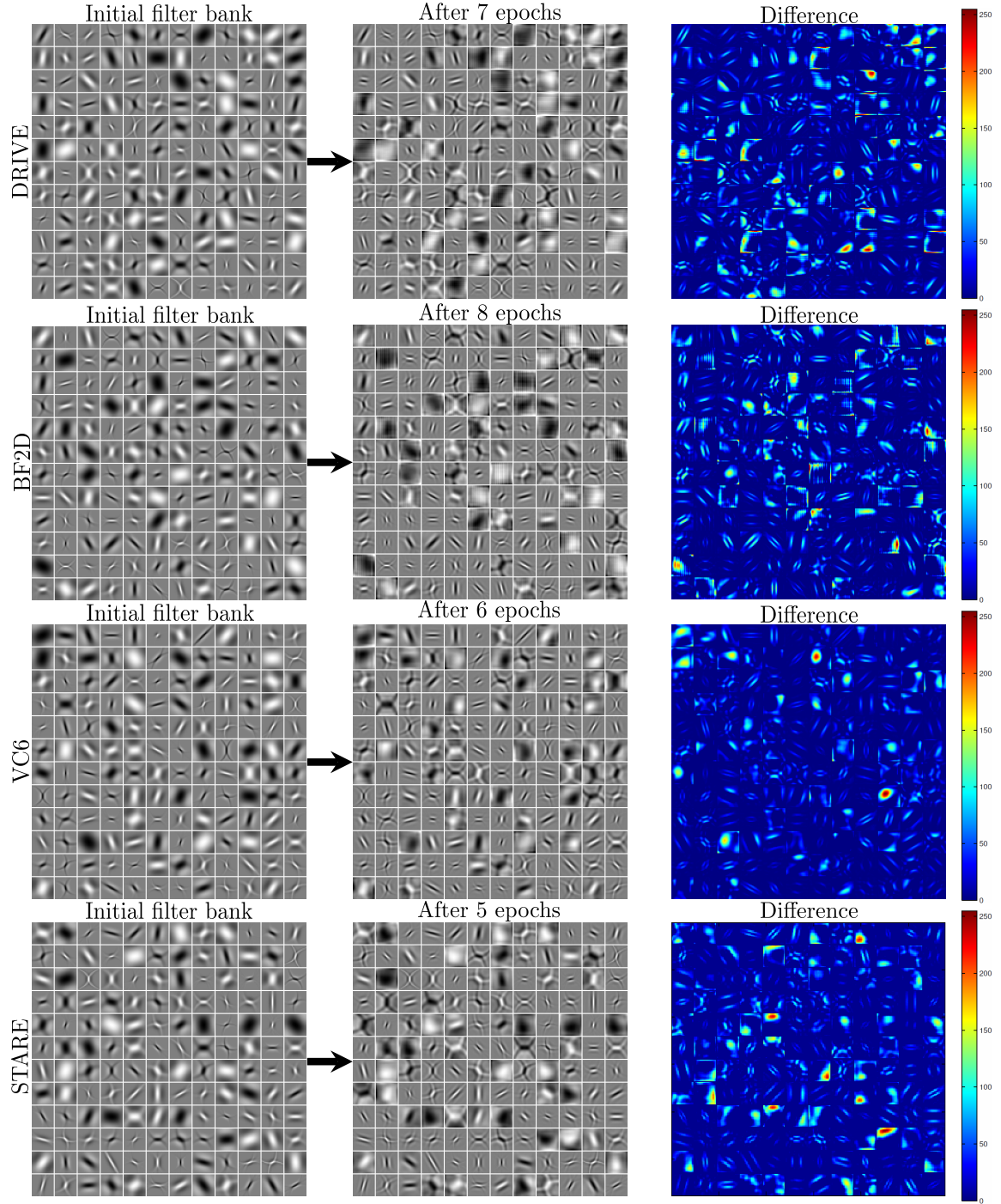


Figure 6.7 **Visualisation of a CSC-refined SCIRD-TS filter bank.** SCIRD-TS filter banks obtained after the fast warm-start strategy (first column), refinement by CSC (second column) and difference (third column) for DRIVE, BF2D, VC6 and STARE (refer to Table 6.1 for time to convergence). Some of the original filters are unchanged, while most of the others are only modified in length or width.

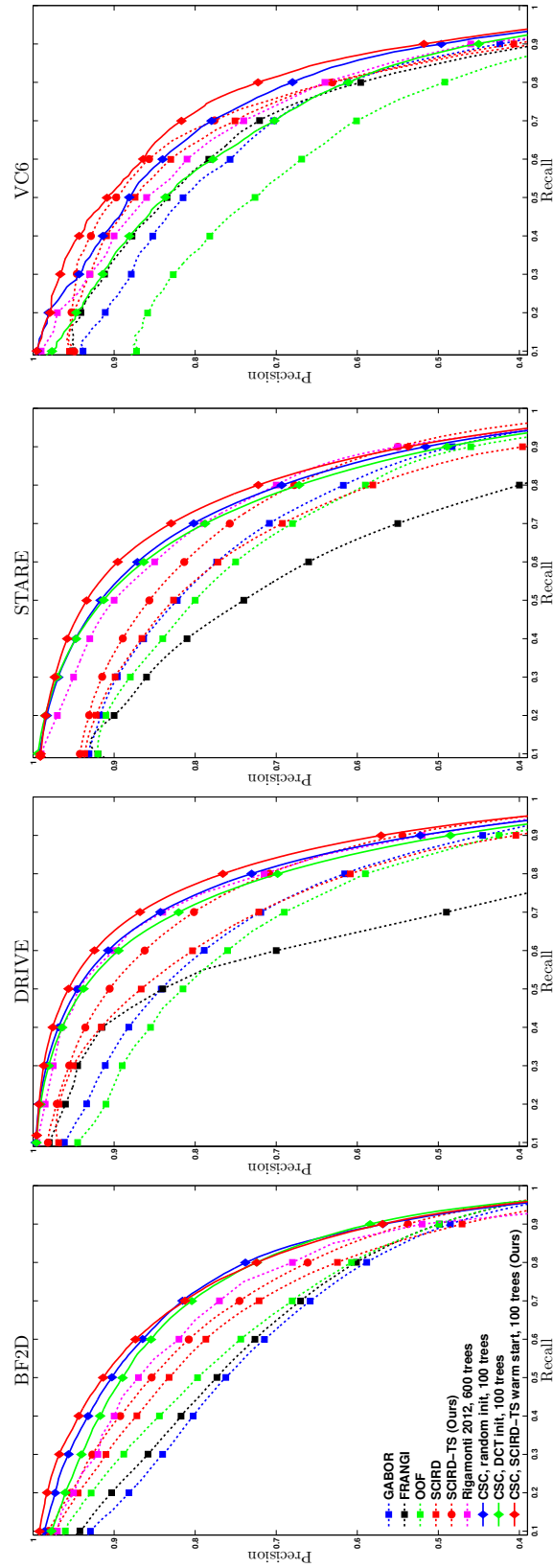


Figure 6.8 **Experiments: segmentation.** Performance evaluation in terms of precision-recall curves for pixel-level segmentation. Notice that I employed a RF with only 100 trees, compared to the method proposed by Rigamonti et al.[111] in which 600 trees were used, hence slower at testing time.

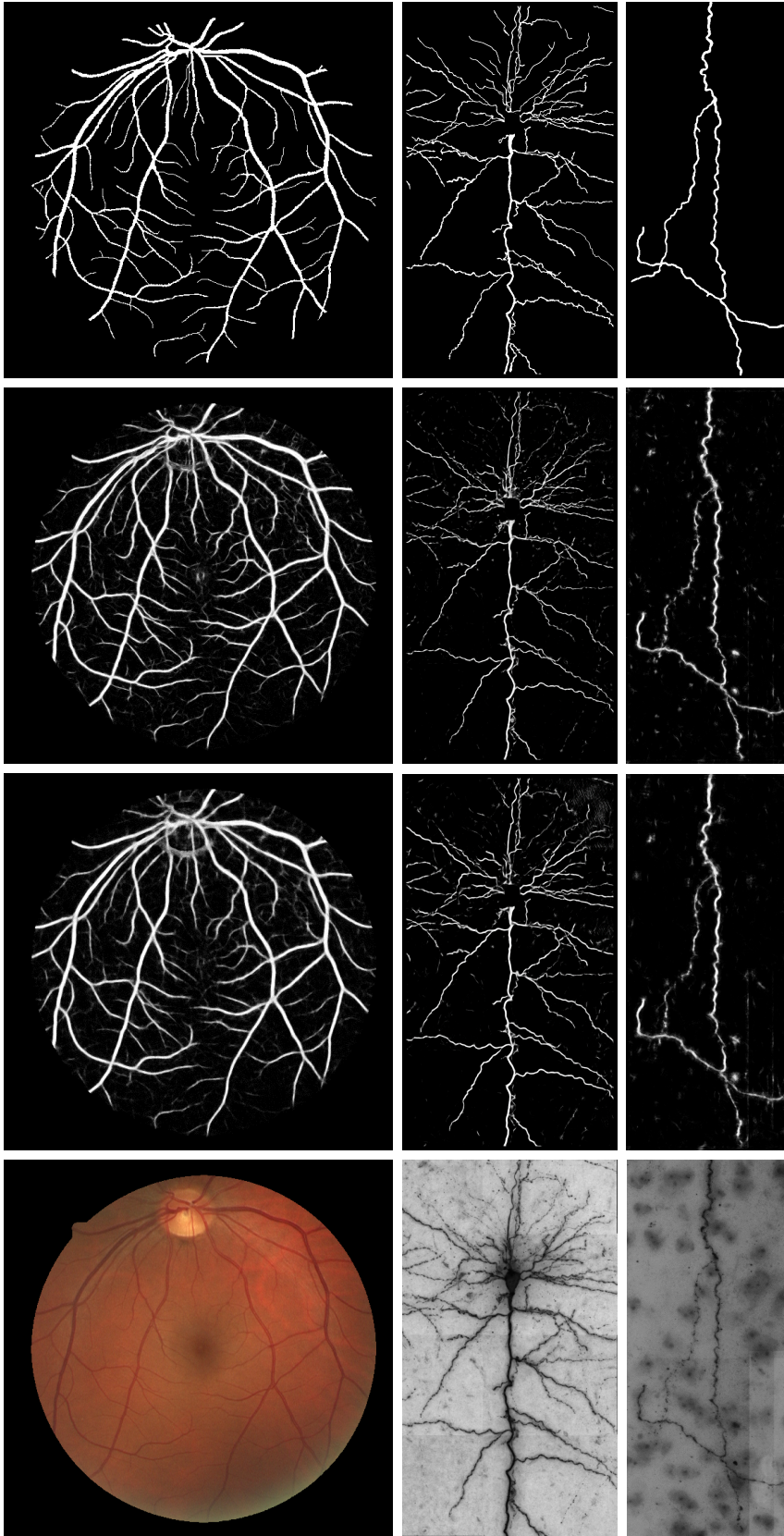


Figure 6.9 **Experiments: segmentation.** Probability maps computed on images from DRIVE (first row), BF2D (second row) and VC6 (third row). For each row, from left to right, we report original image, result of the best performing baseline (i.e. “CSC, random init.”), proposed method’s result and ground truth.

Chapter 7

Conclusions, Discussion and Future Work

7.1 Introduction

In this chapter I review and summarise the work presented in this thesis. I also discuss the limitations of the proposed tortuosity estimation system and suggest potential solutions/extensions to explore in the future.

7.2 Summary of the thesis

Several studies have reported correlations between various diseases and the tortuosity of anatomical curvilinear structures. Such studies are often based on time-consuming, manual annotations and subjective, visual assessments, thus reducing repeatability and inter-observer agreement. I have addressed these problems by proposing a fully automated framework for image-level tortuosity estimation with application to corneal nerve fibres in IVCN.

The proposed system includes two modules: segmentation and tortuosity estimation. The former relies on a hybrid segmentation method combining an appearance model, based on a scale and curvature invariant ridge detector (SCIRD), with a context model, including multi-range learned context filters. The latter is based on a novel tortuosity estimation

paradigm, which identifies the most discriminative tortuosity features and their combination, from the pool of multi-scale tortuosity measures.

I have validated each module of the system separately and compared their performance with state-of-the-art algorithms performing the same task (i.e. segmentation or tortuosity estimation). Then, I have assessed the tortuosity estimation performance of the whole system and compared with that of baseline methods and cornea specialists. Finally, I have assessed the impact of the segmentation module on the tortuosity estimation performance and compared it with manual and traditional segmentation approaches.

Experiments on tortuosity estimation (i.e. the target application) have been carried out on 140 images from healthy subjects and subjects with different pathologies. This data set represents the largest ever adopted to validate a fully automated system for corneal nerve fibre tortuosity estimation. To the best of my knowledge, the largest previous data set included 30 images only. Importantly, I assessed tortuosity estimation performance with 4 levels, whereas to the best of my knowledge, previous studies discriminated among 3 classes, at most.

Experimental results show that the proposed segmentation module for tortuous structures outperforms conventional methods based on the “locally-straight” assumption and learned appearance filters. The improvement in terms of segmentation is immediately transferred to tortuosity estimation performance. The tortuosity estimation module, based on multi-scale spline-based curvature estimation, performs considerably better than state-of-the-art *single*-index algorithms and addresses the main drawback of the previously proposed multi-scale-multi-window approach, i.e. speed. In fact, the proposed solution is orders of magnitude faster than the previous one, and achieves better performance. This speed gain, combined with the hybrid segmentation solution, makes the system fast: 30s are required to analyse an IVCN image (384×384 pixels) using a standard computer with MATLAB. The comparison with three experienced observers who annotated the images independently shows that the proposed system matches or exceeds their performance (Section 5.2.4). These pilot-level experiments indicate the feasibility of large screening programs, subject of course to further validation with much larger data sets.

Although my target application was corneal nerve fibre tortuosity estimation, I have carried out experiments on other structures such as retinal blood vessels and neurites with the aim of making the proposed system versatile. To this end, I have provided a detailed discussion about our system parametrisation and guidelines to set its parameters.

The following sections summarise the detailed contents of the thesis.

7.2.1 Curvilinear structure modelling and segmentation

Chapter 4 focussed on the segmentation module of the fully automated system for tortuosity estimation. A hybrid approach to curvilinear structure segmentation was adopted, with the aim of making the solution suitable to analyse large volumes of images efficiently. Specifically, it combined a curved-support *model-based* ridge detector (SCIRD) with multi-range learned context filters.

Modelling a tortuous curvilinear structure by means of a curved-support Gaussian, I derived a ridge detector measuring the contrast between the part inside and outside the ridge, extending the basic idea of traditional (i.e. locally-straight) ridge detectors. Experiments carried out on 3 challenging and diverse data sets showed that SCIRD outperforms state-of-the-art HCFs.

To compensate for the inevitable modelling limitations of HCFs I proposed to combine SCIRD with learned context filters. This has multiple advantages, including the efficient modelling of inter-object relationships without employing expensive (in terms of running time) multi-layer architectures, such as auto-context. First, I showed that learning context filters is more discriminative than learning appearance ones, as recently proposed. Second, I have discussed the main limitations of adopting single-range context filters and proposed an efficient way of overcoming them by learning multi-range context filters. Experimental results on the target data set (IVCM140) showed that the combination of SCIRD with multi-range context filters outperforms a hybrid approach and several state-of-the-art HCFs.

As discussed in Chapter 5, the improvement in terms of segmentation performance (compared to traditional segmentation approaches based on locally-straight assumption)

transfers immediately to tortuosity estimation performance, hence motivating the need for the proposed algorithms.

7.2.2 Tortuosity estimation

Chapter 5 discussed the tortuosity estimation module of the proposed fully automated tortuosity estimation system. To overcome the limitation of previous work on automated tortuosity estimation (Chapter 2.3), a machine learning approach was adopted. First, a multi-scale tortuosity representation of each curvilinear structure within an image is adopted. Second, image-level tortuosity representations are obtained by weighted averaging (where fibres' lengths were used as weights). Third, a wrapper-based feature selection approach was employed to automatically identify the best combination of tortuosity measures and their relative weight. This had the important advantage of making the proposed system versatile and capable of adapting itself to different curvilinear structures and tortuosity characteristics which may vary among pathologies. Finally, a novel visualisation tool for tortuosity interpretation has been introduced and its advantages have been discussed.

Experiments on the target data set (IVCM140) showed that the proposed tortuosity estimation approach outperforms state-of-the-art tortuosity indices, based on a postulated tortuosity definition, hence justifying the need for a versatile solution such as the one proposed. Remarkably, the proposed fully automated tortuosity estimation system matches and sometimes exceeds the level of performance of cornea specialists when compared against each other. Moreover, the adoption of a spline-based curvature estimation algorithm makes tortuosity estimation fast, as only 30s are required to estimate the tortuosity of a whole image on average.

7.2.3 Improving curvilinear structure modelling and segmentation

Chapter 6 focussed on improving the segmentation module (often the major cause of errors).

First, I noticed the limitation of SCIRD in detecting very thin structures such as retinal blood vessels acquired with fundus camera around the fovea. With the aim of making the proposed system versatile, I proposed and validated a different formulation of SCIRD, SCIRD-TS. Experiments with retinal blood vessels and neurites showed that SCIRD-TS improves considerably the detection of very thin vessels, hence tackling the limitation of SCIRD, without degrading the performance when dealing with wider structures such as neurites acquired at high resolution.

Second, motivated by the recent success of convolutional sparse coding for filter learning with application to curvilinear structure segmentation, I have discussed a novel approach to address its main limitation: slow training time. The acceleration strategy is based on an optimal warm-start strategy which leverages the modelling efforts of the proposed SCIRD-TS HCFs. Experiments with 4 data sets, including retinal blood vessels and neurites showed that the proposed acceleration strategy reduces the time to learn convolutional filter banks considerably compared to traditional initialisation approaches: about 3 hours are required to learn highly discriminative filter banks (on a standard laptop), compared to days needed to learn such filters with a traditional approach. This speed-up would allow learning a large quantity of multi-range context filters that are potentially more discriminative than the ones currently adopted, based on K-means clustering. Although the current segmentation solution seems to provide good segmentation results, the possibility of modelling context with more effective solutions may be needed to achieve specialist-level tortuosity estimation performance with other data sets.

7.3 Contributions

In this thesis I proposed a fully automated system for quantifying the tortuosity of curvilinear structures in medical images. I contributed to the existing literature of both curvilinear structure segmentation and tortuosity estimation; the key contributions can be summarised as follows,

- Novel hand-crafted ridge detectors, SCIRD (Scale and Curvature Invariant Ridge Detector) and SCIRD-TS (SCIRD for very Thin Structures), which are simultaneously rotation, scale, contrast, elongation and, unlike the others, curvature invariant (Sections 4.2 and 6.2).
- Efficient solutions to incorporate single- and multi-range *context* information (i.e. inter-object relationships) for curvilinear structure segmentation, without increasing the computational cost compared with approaches based only on *appearance* information (Sections 4.3 and 4.4).
- A novel approach to accelerate convolutional sparse coding for unsupervised filter learning, based on leveraging carefully designed hand-crafted features (Section 6.3).
- A multi-scale approach to tortuosity quantification, shown to be more suitable to tortuosity estimation than previous, single-scale ones (Section 5.2.1).
- A new paradigm to tortuosity definition based on machine learning, designed to be more versatile than fixed, postulated definitions or indices (Section 5.2).

7.4 Limitations of the proposed tortuosity estimation system and future work

In this section I discuss the main limitations of the proposed system and suggest possible solutions to explore in the future.

7.4.1 Segmentation module

Although the curved-support ridge detector, SCIRD, can be easily extended to 3-D curvilinear structures by modelling them with 3-D curved-support Gaussians, the number of filters to be employed could be very high and pose problems of storage and speed. Three solutions could be explored.

First, ridge detection could be split in two phases: for each sub-volume of the 3-D image, (1) apply local eigenvalue decomposition to obtain the principal directions; (2) apply 2-D SCIRD to a certain amount of 2-D slices containing the principal direction pointing along the structure.

Second, investigate steerability [53] (or an approximation thereof) to reduce the number of 3-D filters to be employed.

Third, similar to the solution proposed in the warm-start strategy (Section 6.3.1) to compress the original SCIRD space into a smaller one, creating a large 3-D filter bank spanning the range of parameters characterising the curvilinear structures under analysis and then approximate this filter bank with a much smaller one. Minimising the reconstruction error would be the first approach to try, but others could be investigated.

7.4.2 Tortuosity estimation module

Although IVCMI40 is, to my best knowledge, the largest ever used data set for image-level corneal nerve fibre tortuosity estimation, it is still small to adopt more complex machine learning solutions compared to the simple MLOR model. Once a much larger data set and image-level annotations become available, an ordinal regressor based on SVM or decision forest could lead to better classification performance.

7.4.3 Interpretation of tortuosity estimates

The use of the tortuosity plane yields a better interpretation of the tortuosity estimation results in various ways. However, the automatic feature selection procedure, running on a different data set including different curvilinear structures and pathologies, could identify a combination of more than 2 tortuosity measures as optimal. If the best combination identified includes more than 3 tortuosity measures, a 3-D volume would not be sufficient to visualise the tortuosity *space*. In that case, a solution could be the use of bar plots, where each bar would be a specific tortuosity measure. Alternatively, the multi-dimensional feature space could be projected onto a 2-D or 3-D space by means of a mapping transformation.

7.4.4 Experiments

Tortuosity has been investigated with qualitative and semi-quantitative approaches for several other structures in the human body, as discussed in Section 1.1. In these cases, the proposed system could be employed to increase repeatability and eliminate subjectivity. Although the system could be adapted to work on different curvilinear structures and acquisition modalities, as discussed in Section 5.2.4, testing tortuosity estimation performance in pilot-level studies would be an important preliminary step. Once the latter have been carried out, the current system could be deployed to analyse large volumes of image data, such as the one included in the UKBiobank repository of fundus camera images [95]. The end goal would be investigating tortuosity as biomarker in several pathologies for which no study has been carried out so far.

Moreover, since the proposed system offers promising segmentation performance, other morphometric parameters of curvilinear structures could be extracted and investigated; for instance, the density of corneal nerve fibres within IVCN images, currently measured with semi-automated approaches.

Bibliography

- [1] Akbar, S., Jordan, L., Thompson, A. M., and McKenna, S. J. (2015). Tumor localization in tissue microarrays using rotation invariant superpixel pyramids. In *IEEE International Symposium on Biomedical Imaging (ISBI 2015)*, pages 1292–1295.
- [2] Al-Diri, B. and Hunter, A. (2009). Automated measurements of retinal bifurcations. In *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*, pages 205–208. Springer.
- [3] Al-Diri, B., Hunter, A., and Steel, D. (2009). An active contour model for segmenting and measuring retinal vessels. *IEEE Transactions on Medical Imaging*, 28:1488–1497.
- [4] Annunziata, R., Garzelli, A., Ballerini, L., Mecocci, A., and Trucco, E. (2015a). Leveraging multiscale Hessian-based enhancement with a novel exudate inpainting technique for retinal vessel segmentation. *IEEE Journal of Biomedical and Health Informatics* (in press).
- [5] Annunziata, R., Kheirkhah, A., Aggarwal, S., Cavalcanti, B. M., Hamrah, P., and Trucco, E. (2014). Tortuosity classification of corneal nerves images using a multiple-scale-multiple-window approach. In *Proceedings of the Ophthalmic Medical Image Analysis (OMIA) First International Workshop, MICCAI 2014.*, pages p–113. Chen X, Garvin MK, Liu JJ editors.
- [6] Annunziata, R., Kheirkhah, A., Hamrah, P., and Trucco, E. (2015b). Boosting hand-crafted features for curvilinear structure segmentation by learning context filters. In *Medical Image Computing and Computer-Assisted Interventions (MICCAI 2015)*, pages 596–603. LNCS.
- [7] Annunziata, R., Kheirkhah, A., Hamrah, P., and Trucco, E. (2015c). Scale and curvature invariant ridge detector for tortuous and fragmented structures. In *Medical Image Computing and Computer-Assisted Interventions (MICCAI 2015)*, pages 588–595. LNCS.
- [8] Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.
- [9] Azemin, M. Z. C., Kumar, D. K., Wong, T. Y., Kawasaki, R., Mitchell, P., and Wang, J. J. (2011). Robust methodology for fractal analysis of the retinal vasculature. *IEEE Transactions on Medical Imaging*, 30(2):243–250.

- [10] Azzopardi, G., Strisciuglio, N., Vento, M., and Petkov, N. (2015). Trainable cosfire filters for vessel delineation with application to retinal images. *Medical Image Analysis*, 19(1):46–57.
- [11] Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. (2011). Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, 5.
- [12] Bao, C., Ji, H., Quan, Y., and Shen, Z. (2015). Dictionary learning for sparse coding: Algorithms and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (in Press)*.
- [13] Bao, C., Quan, Y., and Ji, H. (2014). A convergent incoherent dictionary learning algorithm for sparse coding. In *European Conference on Computer Vision—ECCV 2014*, pages 302–316. Springer.
- [14] Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.
- [15] Becker, C., Rigamonti, R., Lepetit, V., and Fua, P. (2013). Supervised feature learning for curvilinear structure segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*, pages 526–533. Springer.
- [16] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127.
- [17] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.
- [18] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- [19] Bogunović, H., Pozo, J. M., Cárdenes, R., Villa-Uriol, M. C., Blanc, R., Pötin, M., and Frangi, A. F. (2012). Automated landmarking and geometric characterization of the carotid siphon. *Medical Image Analysis*, 16(4):889–903.
- [20] Boyd, K., Eng, K. H., and Page, C. D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases*, pages 451–466. Springer.
- [21] Brebisson, A. and Montana, G. (2015). Deep neural networks for anatomical brain segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2015)*, pages 20–28.
- [22] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- [23] Bribiesca, E. (2013). A measure of tortuosity based on chain coding. *Pattern Recognition*, 46(3):716 – 724.
- [24] Bristow, H., Eriksson, A., and Lucey, S. (2013). Fast convolutional sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 391–398.

- [25] Brown, K. M., Barrionuevo, G., Canty, A. J., De Paola, V., Hirsch, J. A., Jefferis, G. S., Lu, J., Snippe, M., Sugihara, I., and Ascoli, G. A. (2011). The DIADEM data sets: representative light microscopy images of neuronal morphology to advance automation of digital reconstructions. *Neuroinformatics*, 9(2-3):143–157.
- [26] Budai, A., Hornegger, J., and Michelson, G. (2009). Multiscale approach for blood vessel segmentation on retinal fundus images. *Investigative Ophthalmology & Visual Science*, 50(13):325–325.
- [27] Bullitt, E., Ewend, M. G., Aylward, S., Lin, W., Gerig, G., Joshi, S., Jung, I., Muller, K., and Smith, J. K. (2004). Abnormal vessel tortuosity as a marker of treatment response of malignant gliomas: preliminary report. *Technology in Cancer Research & Treatment*, 3(6):577–584.
- [28] Bullitt, E., Gerig, G., Pizer, S. M., Lin, W., and Aylward, S. R. (2003). Measuring tortuosity of the intracerebral vasculature from MRA images. *IEEE Transactions on Medical Imaging*, 22(9):1163–1171.
- [29] Bullitt, E., Muller, K. E., Jung, I., Lin, W., and Aylward, S. (2005). Analyzing attributes of vessel populations. *Medical Image Analysis*, 9(1):39–49.
- [30] Cairney, J. (1924). Tortuosity of the cervical segment of the internal carotid artery. *Journal of Anatomy*, 1(59):87–96.
- [31] Capowski, J. J., Kylstra, J. A., and Freedman, S. F. (1995). A numeric index based on spatial frequency for the tortuosity of retinal vessels and its application to plus disease in retinopathy of prematurity. *Retina*, 15(6):490–500.
- [32] Chalasani, R., Principe, J. C., and Ramakrishnan, N. (2013). A fast proximal method for convolutional sparse coding. In *International Joint Conference on Neural Networks (IJCNN 2013)*, pages 1–5. IEEE.
- [33] Cheung, C. Y.-l., Zheng, Y., Hsu, W., Lee, M. L., Lau, Q. P., Mitchell, P., Wang, J. J., Klein, R., and Wong, T. Y. (2011). Retinal vascular tortuosity, blood pressure, and cardiovascular risk factors. *Ophthalmology*, 118(5):812–818.
- [34] Ciresan, D., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in Neural Information Processing Systems (NIPS 2012)*, pages 2843–2851.
- [35] Ciresan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. Springer.
- [36] Coates, A. and Ng, A. Y. (2012). Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade*, pages 561–580. Springer.
- [37] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

- [38] Coulston, J., Baigent, A., Selvachandran, H., Jones, S., Torella, F., and Fisher, R. (2014). The impact of endovascular aneurysm repair on aortoiliac tortuosity and its use as a predictor of iliac limb complications. *Journal of Vascular Surgery*, 60(3):585 – 589.
- [39] Criminisi, A. and Shotton, J. (2013). *Decision forests for computer vision and medical image analysis*. Springer.
- [40] Criminisi, A., Shotton, J., and Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227.
- [41] Cruzat, A., Witkin, D., Baniasadi, N., Zheng, L., Ciolino, J. B., Jurkunas, U. V., Chodosh, J., Pavan-Langston, D., Dana, R., and Hamrah, P. (2011). Inflammation and the nervous system: The connection in the cornea in patients with infectious keratitis. *Investigative Ophthalmology & Visual Science*, 52(8):5136–5143.
- [42] Doubal, F. N., Hokke, P. E., and Wardlaw, J. M. (2009). Retinal microvascular abnormalities and stroke: a systematic review. *Journal of Neurology, Neurosurgery & Psychiatry*, 80(2):158–165.
- [43] Dougherty, G. and Varro, J. (2000). A quantitative index for the measurement of the tortuosity of blood vessels. *Medical Engineering & Physics*, 22(8):567–574.
- [44] Dunn, G. (1989). *Design and analysis of reliability studies: The statistical evaluation of measurement errors*. Edward Arnold Publishers.
- [45] Edington, G. (1901). Tortuosity of both internal carotid arteries. *British Medical Journal*, 2(2134):1526–7.
- [46] Edwards, K., Pritchard, N., Vagenas, D., Russell, A., Malik, R. A., and Efron, N. (2014). Standardizing corneal nerve fibre length for nerve tortuosity increases its association with measures of diabetic neuropathy. *Diabetic Medicine*, 31(10):1205–1209.
- [47] Eleid, M. F., Guddeti, R. R., Tweet, M. S., Lerman, A., Singh, M., Best, P. J., Vrtiska, T. J., Prasad, M., Rihal, C. S., Hayes, S. N., et al. (2014). Coronary artery tortuosity in spontaneous coronary artery dissection angiographic characteristics and clinical implications. *Circulation: Cardiovascular Interventions*, 7(5):656–662.
- [48] Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929.
- [49] Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. (1998). Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 1998)*, pages 130–137. Springer.
- [50] Franken, R., el Morabit, A., de Waard, V., Timmermans, J., Scholte, A. J., van den Berg, M. P., Marquering, H., Planken, N. R., Zwinderman, A. H., Mulder, B. J., et al. (2015). Increased aortic tortuosity indicates a more severe aortic phenotype in adults with marfan syndrome. *International Journal of Cardiology*, 194:7–12.

- [51] Ganin, Y. and Lempitsky, V. (2014). N4-fields: Neural network nearest neighbor fields for image transforms. In *Asian Conference on Computer Vision–ACCV 2014*, pages 536–551. Springer.
- [52] Ghadiri, F., Pourreza, H., and Banaee, T. (2012). A novel automatic method for vessel tortuosity evaluation. In *International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 56–59. IEEE.
- [53] González, G., Aguet, F., Fleuret, F., Unser, M., and Fua, P. (2009). Steerable features for statistical 3d dendrite detection. In *MICCAI*.
- [54] Grisan, E., Foracchia, M., and Ruggeri, A. (2008). A novel method for the automatic grading of retinal vessel tortuosity. *IEEE Transactions on Medical Imaging*, 27(3):310–319.
- [55] Gu, L. and Cheng, L. (2015). Learning to boost filamentary structure segmentation. In *IEEE International Conference on Computer Vision – ICCV 2015*, pages 639–647.
- [56] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.
- [57] Hamrah, P., Cruzat, A., Dastjerdi, M. H., Prüss, H., Zheng, L., Shahatit, B. M., Bayhan, H. A., Dana, R., and Pavan-Langston, D. (2013). Unilateral herpes zoster ophthalmicus results in bilateral corneal nerve alteration: An in vivo confocal microscopy study. *Ophthalmology*, 120(1):40 – 47.
- [58] Hamrah, P., Cruzat, A., Dastjerdi, M. H., Zheng, L., Shahatit, B. M., Bayhan, H. A., Dana, R., and Pavan-Langston, D. (2010). Corneal sensation and subbasal nerve alterations in patients with herpes simplex keratitis: An in vivo confocal microscopy study. *Ophthalmology*, 117(10):1930 – 1936.
- [59] Han, H.-C. (2009). Blood vessel buckling within soft surrounding tissue generates tortuosity. *Journal of Biomechanics*, 42(16):2797–2801.
- [60] Hannink, J., Duits, R., and Bekkers, E. (2014). Crossing-preserving multi-scale vesselness. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pages 603–610. Springer.
- [61] Hart, W. E., Goldbaum, M., Côté, B., Kube, P., and Nelson, M. R. (1999). Measurement and classification of retinal vascular tortuosity. *International Journal of Medical Informatics*, 53(2):239–252.
- [62] Hathout, L. and Do, H. M. (2012). Vascular tortuosity: a mathematical modeling perspective. *The Journal of Physiological Sciences*, 62(2):133–145.
- [63] Heide, F., Heidrich, W., and Wetzstein, G. (2015). Fast and flexible convolutional sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 5135–5143.
- [64] Heneghan, C., Flynn, J., O’Keefe, M., and Cahill, M. (2002). Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis. *Medical Image Analysis*, 6(4):407–429.

- [65] Honnorat, N., Vaillant, R., and Paragios, N. (2011). Graph-based geometric-iconic guide-wire tracking. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2011*, pages 9–16. Springer.
- [66] Hoover, A., Kouznetsova, V., and Goldbaum, M. (2000). Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3):203–210.
- [67] Ikram, M., Ong, Y., Cheung, C., and Wong, T. (2013). Retinal vascular caliber measurements: Clinical significance, current knowledge and future perspectives. *Ophthalmologica*, 229(3):125–136.
- [68] Jackson, Z. S., Dajnowiec, D., Gotlieb, A. I., and Langille, B. L. (2005). Partial off-loading of longitudinal tension induces arterial tortuosity. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 25(5):957–962.
- [69] Ji, J., Shimony, J., Gao, F., McKinstry, R. C., and Gutmann, D. H. (2013). Optic nerve tortuosity in individuals with neurofibromatosis type 1. *Pediatric Radiology*, 43(10):1336.
- [70] Joshi, V., Reinhardt, J. M., and Abramoff, M. D. (2010). Automated measurement of retinal blood vessel tortuosity. In *SPIE Medical Imaging*, pages 76243A–76243A. International Society for Optics and Photonics.
- [71] Kalitzeos, A. A., Lip, G. Y., and Heitmar, R. (2013). Retinal vessel tortuosity measures and their applications. *Experimental Eye Research*, 106:40–46.
- [72] Kallinikos, P., Berhanu, M., O'Donnell, C., Boulton, A. J., Efron, N., and Malik, R. A. (2004). Corneal nerve tortuosity in diabetic patients with neuropathy. *Investigative Ophthalmology & Visual Science*, 45(2):418–422.
- [73] Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., and Glocker, B. (2015). Multi-scale 3d convolutional neural networks for lesion segmentation in brain mri. *Ischemic Stroke Lesion Segmentation*, page 13.
- [74] Kass-Iliyya, L., Javed, S., Gosal, D., Kobylecki, C., Marshall, A., Petropoulos, I. N., Ponirakis, G., Tavakoli, M., Ferdousi, M., Chaudhuri, K. R., Jeziorska, M., Malik, R. A., and Silverdale, M. A. (2015). Small fiber neuropathy in parkinson's disease: A clinical, pathological and corneal confocal microscopy study. *Parkinsonism & Related Disorders*, 21(12):1454 – 1460.
- [75] Kavukcuoglu, K., Sermanet, P., Boureau, Y.-L., Gregor, K., Mathieu, M., and Cun, Y. L. (2010). Learning convolutional feature hierarchies for visual recognition. In *Advances in Neural Information Processing Systems (NIPS 2010)*, pages 1090–1098.
- [76] Kotschieder, P., Fiterau, M., Criminisi, A., and Buló', S. R. (2015). Deep neural decision forests. In *International Conference on Computer Vision (ICCV 2015)*, pages 1467–1475.
- [77] Koprowski, R., Teper, S., Weglarz, B., Wylęgała, E., Krejca, M., and Wróbel, Z. (2012). Fully automatic algorithm for the analysis of vessels in the angiographic image of the eye fundus. *Biomedical Engineering Online*, 11(35):10–1186.

- [78] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS 2012)*, pages 1097–1105.
- [79] Kurbanyan, K., Hoesl, L., Schrems, W., and Hamrah, P. (2012). Corneal nerve alterations in acute acanthamoeba and fungal keratitis: an in vivo confocal microscopy study. *Eye*, 26(1):126–132.
- [80] Kylstra, J., Wierzbicki, T., Wolbarsht, M. L., Landers III, M., and Stefansson, E. (1986). The relationship between retinal vessel tortuosity, diameter, and transmural pressure. *Graefes's archive for clinical and experimental ophthalmology*, 224(5):477–480.
- [81] Lagali, N., Poletti, E., Patel, D. V., McGhee, C. N. J., Hamrah, P., Kheirikhah, A., Tavakoli, M., Petropoulos, I. N., Malik, R. A., Utheim, T. P., Zhivov, A., Stachs, O., Falke, K., Peschel, S., Guthoff, R., Chao, C., Golebiowski, B., Stapleton, F., and Ruggeri, A. (2015). Focused tortuosity definitions based on expert clinical assessment of corneal subbasal nerves expert assessment of corneal nerve tortuosity. *Investigative Ophthalmology & Visual Science*, 56(9):5102–5109.
- [82] Lam, B., Gao, Y., and Liew, A.-C. (2010). General retinal vessel segmentation using regularization-based multiconcavity modeling. *IEEE Transactions on Medical Imaging*, 29(7):1369–1381.
- [83] Law, M. W. and Chung, A. C. (2008). Three dimensional curvilinear structure detection using optimally oriented flux. In *European Conference on Computer Vision—ECCV 2008*, pages 368–382. Springer.
- [84] Law, M. W. and Chung, A. C. (2010). An oriented flux symmetry based active contour model for three dimensional vessel segmentation. In *European Conference on Computer Vision—ECCV 2010*, pages 720–734. Springer.
- [85] Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., and Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4):1106–1119.
- [86] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- [87] Lee, H., Battle, A., Raina, R., and Ng, A. Y. (2006). Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NIPS 2006)*, pages 801–808.
- [88] Lesage, D., Angelini, E. D., Bloch, I., and Funka-Lea, G. (2009). A review of 3d vessel lumen segmentation techniques: Models, features and extraction schemes. *Medical Image Analysis*, 13(6):819–845.
- [89] Li, Q., Feng, B., Xie, L., Liang, P., Zhang, H., and Wang, T. (2016a). A cross-modality learning approach for vessel segmentation in retinal images. *IEEE Transactions on Medical Imaging*, 35(1):109–118.

- [90] Li, W., Manivannan, S., Zhang, J., Trucco, E., and McKenna, S. J. (2016b). Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks. In *International Symposium on Biomedical Imaging (ISBI 2016)*.
- [91] Lin, J. K. and Dayan, P. (1999). Curved gaussian models with application to the modeling of foreign exchange rates. In *Computational Finance*. MIT Press.
- [92] Lindeberg, T. (1990). Scale-space for discrete signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(3):234–254.
- [93] Lisowska, A., Annunziata, R., Loh, G., Karl, D., and Trucco, E. (2014). An experimental assessment of five indices of retinal vessel tortuosity with the ret-tort public dataset. In *IEEE Engineering in Medicine and Biology Conference (EMBC 2014)*, pages 5414–5417.
- [94] Longmuir, S. Q., Mathews, K. D., Longmuir, R. A., Joshi, V., Olson, R. J., and Abramoff, M. D. (2010). Retinal arterial but not venous tortuosity correlates with facioscapulohumeral muscular dystrophy severity. *Journal of American Association for Pediatric Ophthalmology and Strabismus*, 14(3):240–243.
- [95] MacGillivray, T. J., Cameron, J. R., Zhang, Q., El-Medany, A., Mulholland, C., Sheng, Z., Dhillon, B., Doubal, F. N., Foster, P. J., Trucco, E., and Sudlow, C. (2015). Suitability of uk biobank retinal images for automatic analysis of morphometric properties of the vasculature. *PLoS ONE*, 5(10):e0127914.
- [96] Marin, D., Aquino, A., Gegundez-Arias, M. E., and Bravo, J. M. (2011). A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features. *IEEE Transactions on Medical Imaging*, 30:146–158.
- [97] Martinez-Perez, M. E., Hughes, A. D., Thom, S. A., Bharath, A. A., and Parker, K. H. (2007). Segmentation of blood vessels from red-free and fluorescein retinal images. *Medical Image Analysis*, 11(1):47–61.
- [98] Maude, R. J., Ahmed, B. W., Rahman, A. H., Rahman, R., Majumder, M. I., Menezes, D. B., Sayeed, A. A., Hughes, L., MacGillivray, T. J., Borooah, S., et al. (2014). Retinal changes in visceral leishmaniasis by retinal photography. *BMC Infectious Diseases*, 14(1):527.
- [99] McKenna, S. J., Amaral, T., Akbar, S., Jordan, L., and Thompson, A. (2013). Immunohistochemical analysis of breast tissue microarray images using contextual classifiers. *Journal of Pathology Informatics*.
- [100] Mendonca, A. and Campilho, A. (2006). Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. *IEEE Transactions on Medical Imaging*, 25(9):1200–1213.
- [101] Muraoka, Y., Tsujikawa, A., Kumagai, K., Akagi-Kurashige, Y., Ogino, K., Murakami, T., Miyamoto, K., and Yoshimura, N. (2014). Retinal vessel tortuosity associated with central retinal vein occlusion: An optical coherence tomography study. *Investigative Ophthalmology & Visual Science*, 55(1):134–141.

- [102] Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Le, Q. V., and Ng, A. Y. (2011). On optimization methods for deep learning. In *International Conference on Machine Learning (ICML 2011)*, pages 265–272.
- [103] Oliveira-Soto, L. and Efron, N. (2001). Morphology of corneal nerves using confocal microscopy. *Cornea*, 20(4):374–384.
- [104] Owen, C. G., Newsom, R. S., Rudnicka, A. R., Barman, S. A., Woodward, E. G., and Ellis, T. J. (2008). Diabetes and the tortuosity of vessels of the bulbar conjunctiva. *Ophthalmology*, 115(6):e27–e32.
- [105] Pahuja, N. K., Shetty, R., Nuijts, R. M., Agrawal, A., Ghosh, A., Jayadev, C., and Nagaraja, H. (2016). An in vivo confocal microscopic study of corneal nerve morphology in unilateral keratoconus. *BioMed Research International*, 2016.
- [106] Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3).
- [107] Payan, A. and Montana, G. (2015). Predicting alzheimer’s disease: a neuroimaging study with 3d convolutional neural networks. *arXiv preprint arXiv:1502.02506*.
- [108] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.
- [109] Poletti, E., Grisan, E., and Ruggeri, A. (2012). Image-level tortuosity estimation in wide-field retinal images from infants with retinopathy of prematurity. In *IEEE Engineering in Medicine and Biology Conference (EMBC 2012)*, pages 4958–4961.
- [110] Ricci, E. and Perfetti, R. (2007). Retinal blood vessel segmentation using line operators and support vector classification. *IEEE Transactions on Medical Imaging*, 26(10):1357–1365.
- [111] Rigamonti, R. and Lepetit, V. (2012). Accurate and efficient linear structure segmentation by leveraging ad hoc features with learned filters. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 189–197. Springer.
- [112] Sandow, S. L., Gzik, D. J., and Lee, R. M. (2009). Arterial internal elastic lamina holes: relationship to function? *Journal of anatomy*, 214(2):258–266.
- [113] Santamaría-Pang, A., Colbert, C., Saggau, P., and Kakadiaris, I. A. (2007). Automatic centerline extraction of irregular tubular structures using probability volumes from multiphoton imaging. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007*, pages 486–494. Springer.
- [114] Sasongko, M., Wong, T., Nguyen, T., Cheung, C., Shaw, J., and Wang, J. (2011). Retinal vascular tortuosity in persons with diabetes and diabetic retinopathy. *Diabetologia*, 54(9):2409–2416.
- [115] Sasongko, M. B., Wong, T. Y., Nguyen, T. T., Cheung, C. Y., Shaw, J. E., Kawasaki, R., Lamoureux, E. L., and Wang, J. J. (2015). Retinal vessel tortuosity and its relation to traditional and novel vascular risk markers in persons with diabetes. *Current Eye Research*, pages 1–7.

- [116] Scarpa, F., Zheng, X., Ohashi, Y., and Ruggeri, A. (2011). Automatic evaluation of corneal nerve tortuosity in images from in vivo confocal microscopy. *Investigative Ophthalmology & Visual Science*, 52(9):6404–6408.
- [117] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.
- [118] Sharrett, A. R., Hubbard, L. D., Cooper, L. S., Sorlie, P. D., Brothers, R. J., Nieto, F. J., Pinsky, J. L., and Klein, R. (1999). Retinal arteriolar diameters and elevated blood pressure: The atherosclerosis risk in communities study. *American Journal of Epidemiology*, 150(3):263–270.
- [119] Sironi, A., Lepetit, V., and Fua, P. (2014). Multiscale centerline detection by learning a scale-space distance transform. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, pages 2697–2704.
- [120] Sironi, A., Lepetit, V., and Fua, P. (2015a). Projection onto the manifold of elongated structures for accurate extraction. In *IEEE International Conference on Computer Vision (ICCV 2015)*, pages 316–324.
- [121] Sironi, A., Tekin, B., Rigamonti, R., Lepetit, V., and Fua, P. (2015b). Learning separable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):94–106.
- [122] Sironi, A., Turetken, E., Lepetit, V., and Fua, P. (2015c). Multiscale centerline detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (in Press)*.
- [123] Soares, J. V., Leandro, J. J., Cesar Jr, R. M., Jelinek, H. F., and Cree, M. J. (2006). Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. *IEEE Transactions on Medical Imaging*, 25(9):1214–1222.
- [124] Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.
- [125] Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., and van Ginneken, B. (2004). Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23:501–509.
- [126] Stosic, T. and Stosic, B. D. (2006). Multifractal analysis of human retinal vessels. *IEEE Transactions on Medical Imaging*, 25(8):1101–1107.
- [127] Strimbu, K. and Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463.
- [128] Szlam, A. D., Gregor, K., and Cun, Y. L. (2011). Structured sparse coding via lateral inhibition. In *Advances in Neural Information Processing Systems (NIPS 2011)*, pages 1116–1124.
- [129] Țălu, Ș. and Giovanzana, S. (2012). Image analysis of the normal human retinal vasculature using fractal geometry. *HVM Bioflux*, 4(1):14–18.

- [130] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.
- [131] Trucco, E., Azegrouz, H., and Dhillon, B. (2010). Modeling the tortuosity of retinal vessels: does caliber play a role? *IEEE Transactions on Biomedical Engineering*, 57(9):2239–2247.
- [132] Trucco, E., Ruggeri, A., Karnowski, T., Giancardo, L., Chaum, E., Hubschman, J. P., al Dir, B., Cheung, C. Y., Wong, D., Abràmoff, M., Lim, G., Kumar, D., Burlina, P., Bressler, N. M., Jelinek, H. F., Meriaudeau, F., Quellec, G., MacGillivray, T., and Dhillon, B. (2013). Validating retinal fundus image analysis algorithms: Issues and a proposal validating retinal fundus image analysis algorithms. *Investigative Ophthalmology & Visual Science*, 54(5):3546.
- [133] Tu, Z. and Bai, X. (2010). Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1744–1757.
- [134] Turior, R., Onkaew, D., Uyyanonvara, B., and Chutinantvarodom, P. (2013). Quantification and classification of retinal vessel tortuosity. *SCIENCE ASIA*, 39(3):265–277.
- [135] Wang, J. J., Liew, G., Wong, T. Y., Smith, W., Klein, R., Leeder, S. R., and Mitchell, P. (2006). Retinal vascular calibre and the risk of coronary heart disease-related death. *Heart*, 92(11):1583–1587.
- [136] Wilson, C. M., Cocker, K. D., Moseley, M. J., Paterson, C., Clay, S. T., Schulenburg, W. E., Mills, M. D., Ells, A. L., Parker, K. H., Quinn, G. E., et al. (2008). Computerized analysis of retinal vessel width and tortuosity in premature infants. *Investigative Ophthalmology & Visual Science*, 49(8):3577–3585.
- [137] Wong, T. Y. and Mitchell, P. (2004). Hypertensive retinopathy. *New England Journal of Medicine*, 351(22):2310–2317.
- [138] Zhao, Y., Rada, L., Chen, K., Harding, S., and Zheng, Y. (2015). Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retinal images. *IEEE Transactions on Medical Imaging*, 34(9):1797–1807.

Appendix A

Further Work Carried Out During the Project

A.1 Investigating the biological factors generating tortuosity

In this thesis I have applied a machine learning approach to tortuosity definition, which has the advantage of being versatile and adapt to different curvilinear structures and pathologies. In this section, I discuss the work related to a definition of tortuosity based on its physiological genesis. In particular, I investigate if and what haemodynamic factors are related to tortuosity. This work was carried out in collaboration with Prof. Dr A. Pries, Dr B. Reglin and their research group at Charité - Universitätsmedizin Berlin (DE).

A.1.1 Related work

A well-known reason for vascular tortuosity is increased blood pressure [33, 59, 62, 68, 80]. For this reason, Kylstra et al. [80] modelled the deformation of a blood vessel experimentally, using a latex tube resting on a horizontal surface. Their experiment was mainly focussed on the observation of shape changes of the tube as result of increasing the pressure inside. They observed that the diameter is more sensitive to changes in pressure

when the pressure is below a certain level. Above this level, buckling occurs and tortuosity increases more rapidly than diameter.

Along with pressure, growth was deemed to be another explanation for buckling or tortuosity. In particular, when a vessel has its end-points fixed, growth beyond a certain level would inevitably lead to buckling and tortuosity. Jackson et al. [68] investigated this aspect thoroughly with rabbit carotid arteries and found that tortuosity increased only after very potent growth inputs. Their explanation was that most arteries, for instance, exhibit substantial *in situ* axial stretch of 40% to 60%. So, they first need to off-load this axial strain before buckling. To investigate the consequences of a partial off-loading of longitudinal tension on the shape of vessels they reduced the axial stretch in rabbit carotid arteries from 60% to 30% using interposition grafts. They observed: (1) no normalisation of axial strain within 12 weeks; (2) all the arteries displayed tissue growth and remodelling that caused tortuosity (despite persistent and lower axial stretch); (3) changes in the vessels structure (enlargement of Internal Elastic Lamina or IEL fenestrae¹). Repeating the same experiment with an inhibitor preventing changes in vessels structure they observed no tortuosity, suggesting that structure is strongly related to tortuosity. This important observation led to complex mathematical tortuosity models that take into account the tissue surrounding the vessel wall [59].

Recently, Hathout and Do [62] investigated which shape properties distinguish physiological from abnormal tortuosity. Based on an optimality criterion (i.e. minimising the average curvature per unit length²) they found that tortuous vessels deviating from a sine-generated curve are abnormal.

A.1.2 Materials

A mesentery network³ including 389 vessels (131 arteries, 132 veins, 126 capillaries) was made available by our collaborators at Charité - Universitätsmedizin Berlin (DE). As

¹The IEL is a key layer of the vessel wall [112]. It includes holes (fenestrae) which contribute to the overall stiffness of a vessel.

²This minimises the changes of direction for the blood flowing through the tortuous vessel.

³The mesentery is a fold of membranous tissue that arises from the posterior wall of the peritoneal cavity and attaches to the intestinal tract.

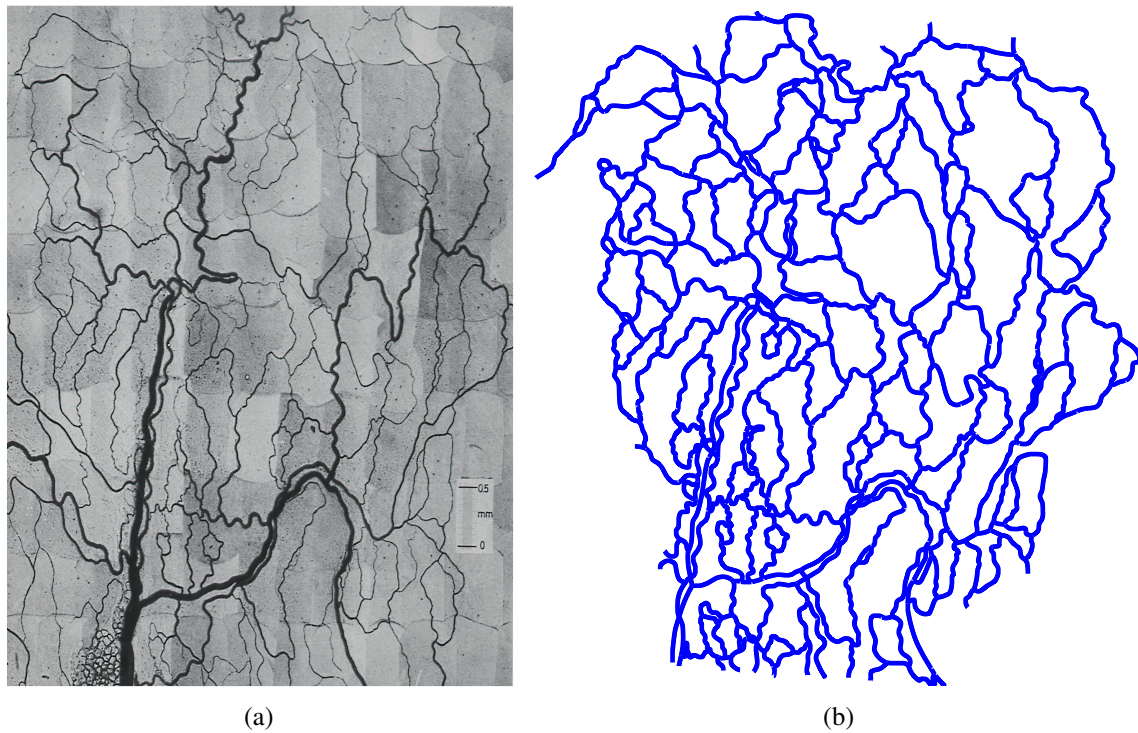


Figure A.1 (a) mesentery network used to investigate biological factors related to tortuosity, (b) full resolution mesentery network obtained by interpolating (using splines) sampling points along each vessel.

can be observed in Figure A.1(a), the network shows a variety of tortuosity characteristics and therefore it is suitable for this study. For each vessel, a number of haemodynamic parameters have been measured or simulated: pressure (Pre), wall shear stress (Wss), diameter (Dia), blood velocity (BV), blood flow (BF), viscosity (Vis), hematocrit (Hem), oxygen partial saturation (PO₂), oxygen saturation (SO₂) and wall thickness (Wt). In addition, I measured local vessel density (Den), as clinical collaborators hypothesize it could be related to tortuosity.

Points along each vessel at high sampling rate have been manually placed by the clinical collaborators and cubic splines were used to reconstruct the whole network (Figure A.1(b)).

Although this data set does not include curvilinear structures such as retinal blood vessels, corneal nerve fibres or neurites (mainly due to data availability) used throughout the thesis, the analysis carried out could give useful insights into the genesis of tortuosity, in general.

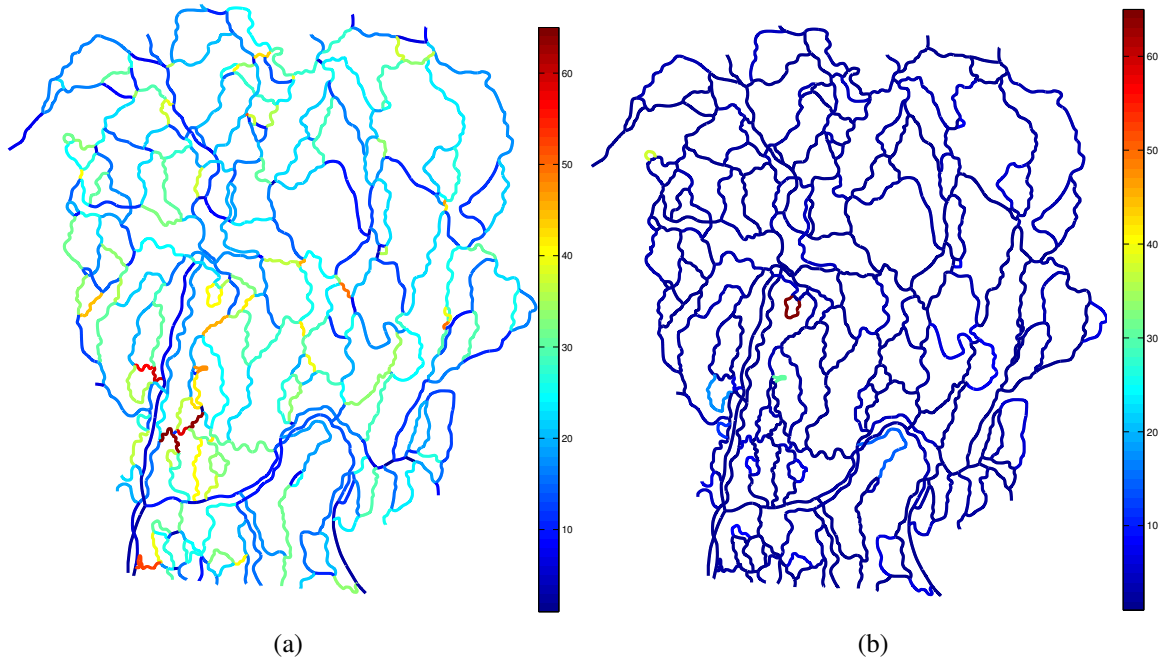


Figure A.2 Tortuosity maps obtained by colour encoding the mean curvature at spatial scale 1 (a) and 10 (b).

A.1.3 Methods

I computed the tortuosity measures described in Section 5.2.1. To set the maximum spatial scale to consider, I plotted tortuosity maps for each measure, as shown by the example in Figure A.2 (tortuosity is encoded with colour).

Density was measured as the ratio of the total vessel length (i.e. the total number of vessel pixels) over the area of a squared neighbourhood⁴. To assign a value to each vessel, the squared neighbourhood was centred on its central point. Vessel width is not taken into account when measuring density. An example of density map is shown in Figure A.3.

Then, I measured correlation with each haemodynamic factor separately and in combination, using a lasso-based regressor [130] to identify the weights automatically.

⁴A 1000×1000 region was used as neighbourhood.

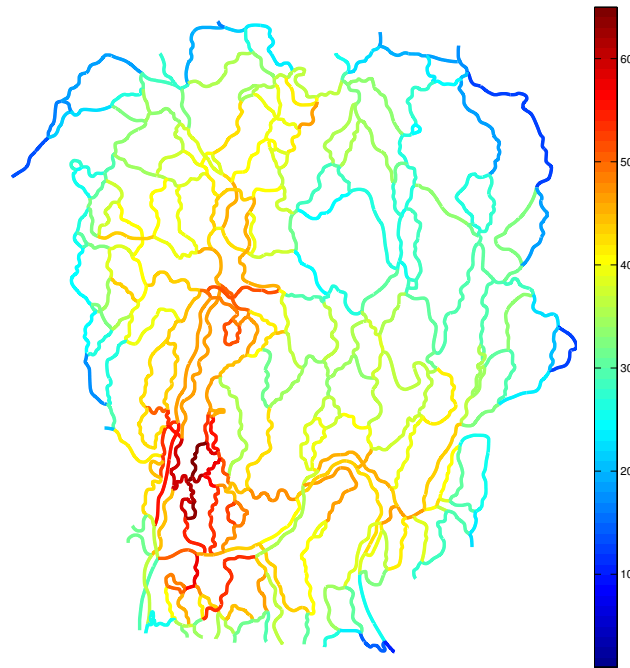


Figure A.3 Density map of the mesentery network in Figure A.1(b).

A.1.4 Experiments and results

Associations have been tested in two ways: (1) finding the combination of tortuosity measures associating best with each individual haemodynamic parameter, and (2) finding the combination of haemodynamic parameters associating best with a tortuosity definition.

Table A.1 reports the correlations related to the experiment (1). These results suggest that tortuosity is related to: pressure, wall shear stress, diameter, blood velocity, viscosity (although the correlation with the hematocrit is not as high⁵), partial but not full oxygen saturation and wall thickness for the arteries; diameter, blood flow, hematocrit (for viscosity correlation is lower) and density for the veins; viscosity (but not hematocrit), partial and full oxygen saturation and density for the capillaries.

It is worth noting that the definition of tortuosity found automatically by the lasso-based regression strategy changes among the haemodynamic parameters and the type of vessel. This is likely due to different structural characteristics of the blood vessels (arteries, veins and capillaries are functionally and structurally different) and potentially due to the differ-

⁵Viscosity and hematocrit levels correlate very well, based on the input from our clinical collaborators.

Table A.1 Experiment (1): measure the correlation of the best combination of tortuosity features with each haemodynamic parameter separately. “NS” replaces correlations whose p-value was greater than 0.01.

	Pre	Wss	Dia	BV	BF	Vis	HEM	PO2	SO2	Den	Wt
Art	0.46	0.48	0.48	0.45	0.37	0.47	0.38	0.47	0.26	NS	0.48
Vei	NS	0.26	0.47	0.24	0.51	0.33	0.41	0.23	0.28	0.49	0.39
Cap	NS	0.38	0.36	NS	0.29	0.50	NS	0.46	0.45	0.41	NS

ent impact of the specific haemodynamic factor on tortuosity. So, they seem to suggest that *adopting a versatile algorithm to re-define tortuosity for different tortuosity estimation problems is needed.*

Another important observation is that correlations between each tortuosity measure and each haemodynamic factor (not reported here for compactness) are typically not very high (i.e. < 0.3), thus suggesting that *tortuosity is caused by a combination of haemodynamic factors.* To investigate this point further, I have run experiments with the lasso-based regression strategy to find the combination of haemodynamic factors correlating best with a tortuosity measure. For example, if we use as tortuosity definition the combination of tortuosity measures giving the highest correlation with a single haemodynamic factor for the arteries (I chose Wss giving 0.48), we find that the best combination of haemodynamic parameters includes all but the density (with different coefficients of the linear combination, of course) and correlation increases to 0.64.

A.1.5 Conclusions

Tortuosity remains very difficult to define objectively. Experiments suggest that *fixed* definitions adopted so far in the literature are not suitable ways to define tortuosity. In this thesis I have proposed a versatile solution to define tortuosity, capable of identifying the best combination of basic tortuosity measures automatically and for each specific problem. Experiments in the previous chapters were based on ground truth provided by expert observers, hence based on perception and subject to inter-observer variability.

In this section I have investigated some of the biological parameters that could potentially cause tortuosity in a mesentery network with the aim of having a better understanding of the basic tortuosity measures to adopt (and potential modifications). Pilot-level experiments seem to confirm that: (1) tortuosity definitions should be versatile and adaptable for different curvilinear structures (e.g. arteries, veins or capillaries); (2) tortuosity does not seem to be caused by a single haemodynamic factor, but it is rather influenced by a combination of them.

Further experiments should be carried out in order to understand how each haemodynamic factor influences tortuosity with the aims of (1) improving tortuosity modelling and transferring these findings to image-level tortuosity estimation; (2) establishing what implications a tortuous structure has for the parameters regulating vital functions in the human body.