Tortuosity Classification of Corneal Nerves Images Using a Multiple-Scale-Multiple-Window Approach

Roberto Annunziata[1] *, Ahmad Kheirkhah[2], Shruti Aggarwal[2], Bernardo M. Cavalcanti[2], Pedram Hamrah[2], and Emanuele Trucco[1]

[1] School of Computing, University of Dundee, Dundee, UK
[2] Department of Ophthalmology, Harvard Medical School, Boston, USA

Abstract. Classify *in vivo* confocal microscopy corneal images by tortuosity is complicated by the presence of variable numbers of fibres of different tortuosity level. Instead of designing a function combining manually selected features into a single coefficient, as done in the literature, we propose a supervised approach which selects automatically the most relevant combination of shape features from a pre-defined dictionary. To our best knowledge, we are the first to consider features at different spatial scales and show experimentally their relevance in tortuosity modelling. Our results, obtained with a set of 100 images and 20 fold cross-validation, suggest that multinomial logistic ordinal regression, trained on consensus ground truth from 3 experts, yields an accuracy indistinguishable, overall, from that of experts when compared against each other.

1 Introduction

Corneal nerves regulate corneal epithelial integrity, proliferation, and wound healing [1]. Recently, quantitative studies using IVCM have led to the development of quantitative methods for morphometric parameters such as width, reflectivity, orientation, branching patterns [1]. We concentrate here on tortuosity. The physiology of nerve fibres tortuosity and its role in pathologies are not yet clear. Recent studies (e.g., [2, 3]) are based on a manual qualitative assessment of corneal nerve tortuosity. Intra-observer and inter-observer variability can be high and unsuitable for an objective and reproducible evaluation [4, 5]. Quantitative measures of vascular tortuosity have been developed for the retinal vasculature as observed in fundus camera images [6–10] and for the 3-D brain vasculature [11]. The distance measure (DM) has been widely used (e.g., [6]), and its limitations have been discussed elsewhere [8]. [10, 11] proposed a different measure in which DM is multiplied by the number of inflection points along vessels. Curvature-based measures are integral functions of curvature estimates

*Corresponding author. Address: Balfour street (Queen Mother Building), Angus DD1 4HN, Dundee, UK. Tel.: +44 (0)1382 386504; fax:+44 (0)1382 385509. Email: r.annunziata@dundee.ac.uk

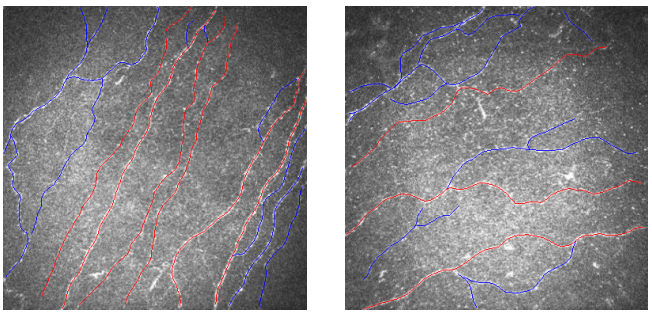
along the vessel's centreline, often weighted sums of absolute or squared curvatures: e.g., [7] compared seven such measures using two classification problems, classifying blood vessel segments as tortuous or non-tortuous, and classifying the whole vessel network. They found the total squared curvature measure to yield the closest results to the ophthalmologist's notion of tortuosity. In comparison, little work exists on the tortuosity of corneal fibres. [5] were the first to propose an objective, semi-automated method for quantifying sub-basal nerve tortuosity. [12] adapted the algorithm in [10] to work with corneal nerve images; the algorithm performed well with a data set of 30 images divided into 3 classes by a single ophthalmologist. We present a novel approach to the estimation of corneal fibre tortuosity, suitable for retinal vessels as well. We do not propose a single formula capturing tortuosity by combining arguably relevant features; instead, we use a supervised-learning approach with feature selection, in which the most discriminative morphometric features (e.g., curvature, number of inflection points) are identified from a dictionary compiled off-line. Moreover, we adopt a multiple-scale approach, so that a trained classifier can automatically learn (1) which feature groups associate with expert's judgement, and (2) what spatial scales are relevant for each feature. We validate our system with a set of 100 corneal nerve images captured using IVCM, and scored independently by the three ophthalmologists authors labelling tortuosity as normal, mild, high and severe. Results show that the accuracy of our method in replicating each observer's judgement is comparable or higher than the best observer hold out (88.89%, 76.67%, 77.22% vs 76.67%, 76.67%, 75%, respectively).

2 Materials

One hundred images of corneal subbasal nerve plexus with different grades of tortuosity, obtained by confocal microscopy were obtained by the clinical authors (removed for anonymity). We observe that similar studies [5, 12] report tests with similar or smaller number of images. The images were chosen so to balance image numbers in each of the 4 tortuosity classes considered. Nerves were traced manually to make the tortuosity estimation independent of the algorithm used to trace them automatically. We concentrate on major nerves in this study (see Fig. 1). Agreement is modest (under 50%), in line with previously reported work [4, 5]. We generate a set of Consensus Ground Truth (CGT) images: at least two observers agreed independently on a class class, which becomes the CGT class for the image. In our dataset, 90 images out of 100 were included into the CGT. We assume all experts equally experienced, which applies to our case.

3 Methods

In essence, we compute image-level features extracted from each fibre using a multiple-scale approach, select automatically the most discriminative ones (best associating with expert judgement), then train a classifier with the selected features. The system is composed by 4 modules: multiple-scale fibre-level features



(a) (b)

Fig. 1. Example of manually traced corneal fibres: (a) low and (b) severe cases. Main fibres are traced in red.

extraction, image-level features combination, feature selection, supervised classification.

3.1 Multiple-scale fibre-level features extraction

Several fibre-level tortuosity features have been proposed in the literature [6–8, 10]. Recent results indicate that the number of inflection points (at which the sign of the curvature changes) and curvature along the fibre or vessel are relevant tortuosity definitions [13], hence we include them in our dictionary of features. Accurate curvature estimation from samples is a well-known tricky task [14–16]. Two main issues are window size selection and accuracy. *Window size selection* is crucial in all algorithms defining a support window, e.g. [15], and errors occur if the window size is not locally adequate. Moreover, the *accuracy* of previously reported methods is not uniformly satisfactory [14–16]. We devised a multiple-window curve fitting algorithm for digital curvature estimation leading to very accurate results and robust to signal variations and noise. The method has these steps:

1. At each pixel, apply a local ellipse fitting *and* line fitting using the smallest window size in a pre-defined range $R = [w_m, w_M]$[3];
2. choose the best fitting function (ellipse or line) based on the sum of squared errors;
3. if ellipse-arc is the best fitting function, compute curvature using analytical derivatives on the estimated ellipse, else consider the curvature zero;
4. repeat 1. - 3. for all windows in R;
5. select the maximum estimated curvature over all windows after median filtering to eliminate small, spurious peaks.

[3] To be determined empirically in order to reject a certain noise level (w_m), to take into account the curvature of the longest fibre's partition between two inflection points (w_M).

We use an ellipse-specific linear least-squares algorithm [17] as the best compromise between accuracy and computational efficiency. As to fibre-level features, we consider: average curvature along the fibre, i.e. k_{mean}; "twistedness" or density of inflection points, d_{ip}; maximum curvature along the fibre k_M. We reckon that density of inflection points, rather than just inflection points number, can yield fairer comparison among fibres of different length; if the number of inflection points per length unit is the same in a long and a short fibres, the two should be considered equally tortuous in terms of inflection points. We also include the maximum curvature in the feature vector to improve discrimination. To capture the full structure of a digital fibre (signal), we use a multiple-scale approach, and build a Gaussian scale-space with discretised Gaussian kernels. In summary, our fibre-level features are: $\{k_{mean}(t)\}$ for $t \in \{1, \ldots, t_M\}$; $\{d_{ip}(t)\}$ for $t \in \{1, \ldots, t_M\}$; $\{k_M(t)\}$ for $t \in \{1, \ldots, t_M\}$; where t_M is the maximum scale t used.

3.2 Image-level features combination

Associating a whole image containing multiple fibres with a tortuosity class is made difficult by the variable number and length of fibres per image, and by the potentially large differences of fibre tortuosity in the same image. In order to solve these problems simultaneously, we use a weighted average of fibre-level features, in which weights are fibre's length at scale t. Denoting the length of the i^{th} fibre as $l_i(t)$ at scale t and the total number of fibres within an image as N_f, our image-level features become the weighted average of:

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

3.3 Feature selection

Feature Selection (FS) highlights at what scale a feature is most relevant for tortuosity assessment, and reveals which features are important at different scales. Here we investigate two main FS approaches: wrappers and embedded methods. Wrappers use the learning machine of interest as a black box to score subsets of feature according to their predictive power. Exhaustive search ("brute force") is guaranteed to find the optimal solution; every other method is sub-optimal. The possibility of using this approach is limited by the number of features and the complexity of the algorithm used to score the selected combination. Since the feature vector we use is made up of a small number of features (i.e. less than 20), a wrapper can be applied to some extent, if paired with a low complexity classification algorithm. When exhaustive search is not feasible, mainly due to the complexity of the classification algorithm, we use a method embedded in the Random Forests (RF) which rank features by importance.

3.4 Supervised classification

As tortuosity classes are naturally ordered, we use two ordinal regression methods: Multinomial Logistic Ordinal Regression (MLOR) and Ordinal SVM (OSVM). Being MLOR relatively simple and fast, we use it in the feature selection wrapper using exhaustive search. Otherwise, we use OSVM [18] (more complex) and RF feature selection (sub-optimal search).

4 Experiments and Results

4.1 Performance measures

We use the standard average Accuracy (Acc), Sensitivity (Se), Specificity (Sp), Positive predictive value (Ppv) and Negative predictive value (Npv). Notice that we use macro-averaging being our dataset balanced (see [19]). To account for ordinal information, we use also Mean Squared Error (MSE) and Mean Absolute Error (MAE).

4.2 Association with consensus GT

In order to assess the ability of the proposed frameworks to replicate the consensus ground truth, we use a double cross-validation (one within the other) on the whole dataset of 90 images (CGT) for each framework. The "external" cross-validation is used for testing, whereas the "internal" is used either to estimate the best model parameters (OSVM) or to select the best combination of features (MLOR) for each unseen testing partition. Given the modest dimension of our dataset, we use a 20-fold testing cross-validation for both MLOR and OSVM to keep the comparison consistent. To assess the best combination of features in the MLOR framework, we let the wrapper procedure score all combinations from $\binom{n_f}{1}$ to $\binom{n_f}{n_{max}}$, where n_f is the total number of features (18, as we consider 6 scales when extracting multiple-scale features) and n_{max} is the maximum number of features in each combination (3 in the best case). Feature selection in the OSVM framework is done by ranking features according to their *importance* estimated by a random forest of 1000 decision trees with a node size of 10. Then, starting from the most important, we sequentially add less important features until n_{max} (4 in the best case) to make combinations which are scored using the out-of-bag error [20]. Once the best combination of features has been estimated, the best cost parameter C_{SVM} of the soft-margin SVM with the perceptron kernel [18] is estimated by the "internal" cross-validation ($\log_2 C_{SVM} = -10, -8, \ldots, 10$). Table 1 shows the comparison between the proposed methods in terms of all performance measures defined in Section 4.1. These results suggest that MLOR better replicates the consensus ground truth. Interestingly, both methods show a MSE slightly higher than MAE, meaning that most of the misclassified images are swapped of just one class, a desirable property in ordinal regression problems.

Table 1. Performance measures of the proposed approaches using CGT as ground truth.

Perfomance measures	**MLOR**	**OSVM**
Acc	84.44%	80.56%
Se	69.77%	61.94%
Sp	89.50%	86.86%
Ppv	69.75%	62.61%
Npv	89.48%	86.84%
MSE	0.3444	0.4222
MAE	0.3222	0.4000

Table 2. Comparison in terms of Accuracy between our best framework based on the MLOR with other observers, when each observer is taken as ground truth. "AK", "PH" and "SA" are the clinical authors. Best results are bolded.

	AK	PH	SA	**MLOR**
AK	100%	76.67%	75%	**88.89%**
PH	**76.67%**	100%	73.89%	**76.67%**
SA	75%	73.89%	100%	**77.22%**

4.3 Association with individual GT

In this experiment, we take the classification given by one ophthalmologist as ground truth, and we compare the performance of our best framework (based on the MLOR and trained using the CGT) with the other two observers. As Table 2 suggests, the agreement of our automatic method is, overall, completely comparable with the agreement among experts.

4.4 Multiple-scale approach justification

To evaluate the effectiveness of a multiple-scale approach, we analyse the best combination of features for each partition in both MLOR and OSVM case. Figure 2 shows the frequency with which combinations are chosen by the wrapper procedure in the MLOR (in the best case as in Section 4.2). The combination $\{K_{mean}(2), K_{mean}(5)\}$ is chosen in 17 out of 20 cases, thus suggesting that weighted mean curvature extracted at multiple scales is very discriminative when using this framework. In the OSVM best case there is not a repeatable best combination, but we still observe that weighted mean curvature at multiple scales is a frequent subset of best features (the subset $\{K_{mean}(3), K_{mean}(4), K_{mean}(5)\}$ is chosen 8 times).

5 Conclusions

Classify IVCM corneal images in terms of tortuosity is made difficult by variable number of fibres and variable level of tortuosity in the same image. We proposed

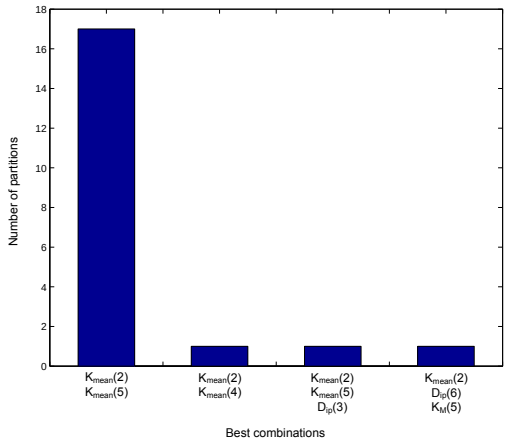


Fig. 2. Frequency of the combinations chosen in the MLOR framework. $\{K_{mean}(2), K_{mean}(5)\}$ is chosen in 17 (out of 20)cases.

a novel supervised approach for tortuosity classification in which discriminative features such as curvature and number of inflection points are automatically selected and combined with the aim of replicating expert's judgement. To the best of our knowledge, we are the the first who take into account scale information which, as our results suggested, is relevant. Experimental results suggest that MLOR, trained using the consensus ground truth, is capable of replicating each observer's predictions as well as the best observer hold out would do or even better, which is the best an automated system can do when tested on ophthalmologist's ground truth. The performances of the proposed frameworks were assessed on unseen images by means of cross-validations.

Acknowledgement

This research was made possible by the EU Marie Curie Initial Training Network (ITN) "REtinal VAscular Modelling, Measurement And Diagnosis" (REVAMMAD), project number 316990. The authors are indebted with Stephen McKenna, Jianguo Zhang and the VAMPIRE group for useful discussions. They are also grateful to Hsuan-Tien Lin for the OSVM code.

References

1. Oliveira-Soto, L., Efron, N.: Morphology of corneal nerves using confocal microscopy. Cornea **20**(4) (2001) 374–384
2. Patel, D., Ku, J., Johnson, R., McGhee, C.: Laser scanning in vivo confocal microscopy and quantitative aesthesiometry reveal decreased corneal innervation and sensation in keratoconus. Eye **23**(3) (2009) 586–592

3. Niederer, R.L., Perumal, D., Sherwin, T., McGhee, C.N.: Laser scanning in vivo confocal microscopy reveals reduced innervation and reduction in cell density in all layers of the keratoconic cornea. Investigative ophthalmology & visual science **49**(7) (2008) 2964–2970
4. De Cillà, S., Ranno, S., Carini, E., Fogagnolo, P., Ceresara, G., Orzalesi, N., Rossetti, L.M.: Corneal subbasal nerves changes in patients with diabetic retinopathy: an in vivo confocal study. Investigative ophthalmology & visual science **50**(11) (2009) 5155–5158
5. Kallinikos, P., Berhanu, M., O'Donnell, C., Boulton, A.J., Efron, N., Malik, R.A.: Corneal nerve tortuosity in diabetic patients with neuropathy. Investigative ophthalmology & visual science **45**(2) (2004) 418–422
6. Heneghan, C., Flynn, J., OKeefe, M., Cahill, M.: Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis. Medical image analysis **6**(4) (2002) 407–429
7. Hart, W.E., Goldbaum, M., Côté, B., Kube, P., Nelson, M.R.: Measurement and classification of retinal vascular tortuosity. International journal of medical informatics **53**(2) (1999) 239–252
8. Trucco, E., Azegrouz, H., Dhillon, B.: Modeling the tortuosity of retinal vessels: does caliber play a role? Biomedical Engineering, IEEE Transactions on **57**(9) (2010) 2239–2247
9. Turior, R., Onkaew, D., Uyyanonvara, B., Chutinantvarodom, P.: Quantification and classification of retinal vessel tortuosity. SCIENCEASIA **39**(3) (2013) 265–277
10. Grisan, E., Foracchia, M., Ruggeri, A.: A novel method for the automatic grading of retinal vessel tortuosity. IEEE Trans. Med. Imaging **27**(3) (2008) 310–319
11. Bullitt, E., Gerig, G., Pizer, S.M., Lin, W., Aylward, S.R.: Measuring tortuosity of the intracerebral vasculature from mra images. Medical Imaging, IEEE Transactions on **22**(9) (2003) 1163–1171
12. Scarpa, F., Zheng, X., Ohashi, Y., Ruggeri, A.: Automatic evaluation of corneal nerve tortuosity in images from in vivo confocal microscopy. Investigative ophthalmology & visual science **52**(9) (2011) 6404–6408
13. Lisowska, A., Annunziata, R., Loh, G., Karl, D., Trucco, E.: An experimental assessment of five indices of retinal vessel tortuosity with the RET-TORT public dataset. In: Engineering in Medicine and Biology Society, 2014. Proceedings of the 36th Annual International Conference of the IEEE. (2014)
14. Coeurjolly, D., Miguet, S., Tougne, L.: Discrete curvature based on osculating circle estimation. In: Visual Form 2001. Springer (2001) 303–312
15. Matas, J., Shao, Z., Kittler, J.: Estimation of curvature and tangent direction by median filtered differencing. In: Image Analysis and Processing, Springer (1995) 83–88
16. Worring, M., Smeulders, A.W.M.: Digitized circular arcs: characterization and parameter estimation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **17**(6) (1995) 587–598
17. Fitzgibbon, A., Pilu, M., Fisher, R.B.: Direct least square fitting of ellipses. Pattern Analysis and Machine Intelligence, IEEE Transactions on **21**(5) (1999) 476–480
18. Li, L., Lin, H.T.: Ordinal regression by extended binary classification. Advances in neural information processing systems **19** (2007) 865
19. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4) (2009) 427–437
20. Breiman, L.: Random forests. Machine learning **45**(1) (2001) 5–32