Sub-category Classifiers for Multiple-instance Learning and Its Application to Retinal Nerve Fiber Layer Visibility Classification

Siyamalan Manivannan[1], Caroline Cobb[2], Stephen Burgess[2], Emanuele Trucco[1]

[1] CVIP, School of Science and Engineering (Computing), University of Dundee, UK
[2] Department of Ophthalmology, NHS Ninewells, Dundee, UK

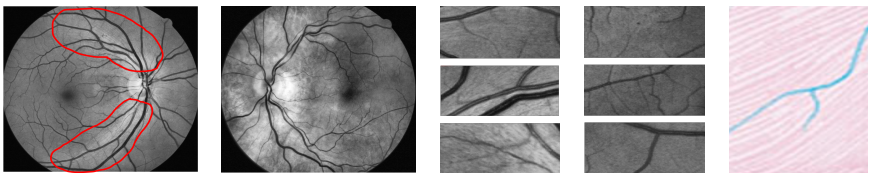
Abstract. We propose a novel multiple instance learning method to assess the visibility (visible/not visible) of the retinal nerve fiber layer (RNFL) in fundus camera images. Using only image-level labels, our approach learns to classify the images as well as to localize the RNFL visible regions. We transform the original feature space to a discriminative subspace, and learn a region-level classifier in that subspace. We propose a margin-based loss function to jointly learn this subspace and the region-level classifier. Experiments with a RNFL dataset containing 576 images annotated by two experienced ophthalmologists give an agreement (kappa values) of 0.65 and 0.58 respectively, with an inter-annotator agreement of 0.62. Note that our system gives higher agreements with the more experienced annotator. Comparative tests with three public datasets (MESSIDOR and DR for diabetic retinopathy, UCSB for breast cancer) show improved performance over the state-of-the-art.

1 Introduction

This paper introduces an automatic system assessing the visibility and location of the RNFL in fundus camera (FC) images from image-level labels. The optic nerve transmits visual information from the retina to the brain. It connects to the retina in the optic disc, and its expansion form the RNFL, the innermost retinal layer. The RNFL has been recently considered as a potential biomarker for dementia [1], by assessing its thickness in optical confocal tomography (OCT) images. However, screening of high numbers of patients would be enabled if the RNFL could be assessed with FC, still much more common than OCT for retinal inspection, and increasingly part of routine optometry checks.

Very little work exists on RNFL-related studies with FC images on studying associations with dementia [2]. This is contrast with RNFL analysis via OCT, supported by a rich literature [1]. The RNFL is not always visible in FC images, and its visibility itself has been posited as a biomarker for neurodegenerative conditions. This motivates our work, part of a larger project on multi-modal retina-brain biomarkers for dementia [3].

[3]EPSRC grant EP/M005976/1



(a) (b) (c) (d) (e)

Fig. 1: RNFL visibility in the green channel: (a) an image with visible RNFL (the marked region indicates its visibility), (b) an image with invisible RNFL, (c) examples of RNFL-visible regions, (d) examples of RNFL-invisible regions, (e) a synthetic image showing RNFL (pink) and blood vessels (blue).

We report an automatic system to identify FC images with visible RNFL regions and simultaneously localize visible regions. A crucial challenge is obtaining ground truth annotations of visible RNFL regions from clinicians, notoriously a difficult and time-consuming process. We take therefore a Multiple Instance Learning (MIL) approach, requiring only image-level labels (RNFL visible/invisible), which can be generated much more efficiently. In MIL, images are regarded as *bags*, and image regions as *instances*.

Visible RNFL regions have significant intra-class variations, and can be difficult to distinguish from RNFL-invisible regions. To address this, we embed the instances in a discriminative subspace defined by the outputs of a set of subcategory classifiers. An instance-level (IL) classifier is then learned in that subspace by maximizing the margin between positive and negative bags. A margin-based loss is proposed to learn the IL and the subcategory classifiers jointly.

Our two main contributions are the following.

1. To our best knowledge, we address a new problem with significant impact potential for biomarker discovery, i.e. classifying FC images as RNFL-visible/invisible, including region localization.
2. We improve experimental performance compared to state-of-the-art MIL systems by proposing a novel MIL approach with a novel margin-based loss (instead of the cross-entropy loss commonly used in comparable MIL systems).

The differences between our and recent, comparable work are captured in Section 2 after a concise discussion of related work.

We evaluated our approach on a local dataset ("RNFL") of 576 FC images, and with three public datasets (MESSIDOR [3] and DR [4] for diabetic retinopathy, UCSB [5] for breast cancer). Table 1 summarizes the datasets and the experimental settings used. The images (green channel) in our RNFL dataset were annotated (image-level annotations) independently by two experienced ophthalmologists (A1 and A2, A1 the more experienced). Overall, they agreed $\simeq 83\%$ of the time ($\mathcal{P} \simeq 83\%$) with a kappa value of $\mathcal{K} \simeq 0.62$. Our experiments suggest that our system highly agrees with A1 than A2 (system agreement with A1, $\mathcal{P} \simeq 84\%$ with $\mathcal{K} \simeq 0.65$ and A2, $\mathcal{P} \simeq 82\%$ with $\mathcal{K} \simeq 0.58$). Our approach also improves the state-of-the-art results on the public datasets (see Table 2).

2 Related Work

MIL approaches can be divided in two broad classes, (1) *instance-level* (IL) and (2) *bag-level* (BL). In both cases a classifier is trained to separate positive from negative bags using a loss function defined at the bag-level. **IL approaches:** the classifier is trained to classify *instances*, obtaining IL predictions. Here, BL predictions are usually obtained by aggregating IL decisions, e.g. MI-SVM [6], MCIL [7]. **BL approaches:** a classifier is trained to classify *bags*. Usually a feature representation is computed for each bag from its instances, then used to learn a supervised classifier. As this is trained at the BL, IL predictions cannot be obtained directly; e.g. JC^{2}MIL [8], and RMC-MIL [9].

The original feature space may not be discriminative. Hence *embedding-based* (EB) approaches try to embed the instances in a discriminative space [8–10]. The bag representation computed in this space is used to learn a BL classifier.

MIL approaches have also been explored within the recent, successful Convolutional Neural Networks (CNN) paradigm for visual recognition [11]. Here, a MIL pooling layer is introduced at the end of the deep network architecture to aggregate (pool) IL predictions and compute the BL ones.

Our approach is an EB approach; but it learns an IL classifier instead of the BL one learned in [8, 9]. Therefore it can provide both IL and BL predictions. CNN+MIL [11], EB approaches [8,9,12] as well as other approaches [7] minimize cross-entropy loss. However, recent results suggest that margin-based loss is better than the cross-entropy loss for classification problems [13]. Considering this, we propose a novel soft-margin loss where the bags which violate the margin are penalized, and show improved performance over the cross-entropy loss.

3 Method

3.1 Motivation and the Overview of the Method

Most MIL approaches do not make explicit assumptions about the inter or intra-class variations of the positive and negative bags (e.g. [6, 14]). However, with high intra-class variation and low inter-class distinction these approaches may not perform well. This is the case for our RNFL dataset: the visible RNFL regions have a high intra-class variations, and they are difficult to distinguish from RNFL-invisible regions (Fig. 1). To overcome this, we assume there exists a set of discriminative sub-categories, and learn a set of classifiers for them. These sub-categories, for instance, may capture different variations (or visual appearance) of the RNFL. Each classifier in this pool is learned specifically to separate a particular sub-category from others. Each instance is thus transformed from its original feature space to a discriminative subspace defined by the output of these classifiers. An IL classifier is then learned in this space by maximizing the margin between the positive and the negative bags. For each bag, the BL prediction is obtained by aggregating (pooling) the decisions of its instances. An overview of the proposed approach is illustrated in Fig. 2.

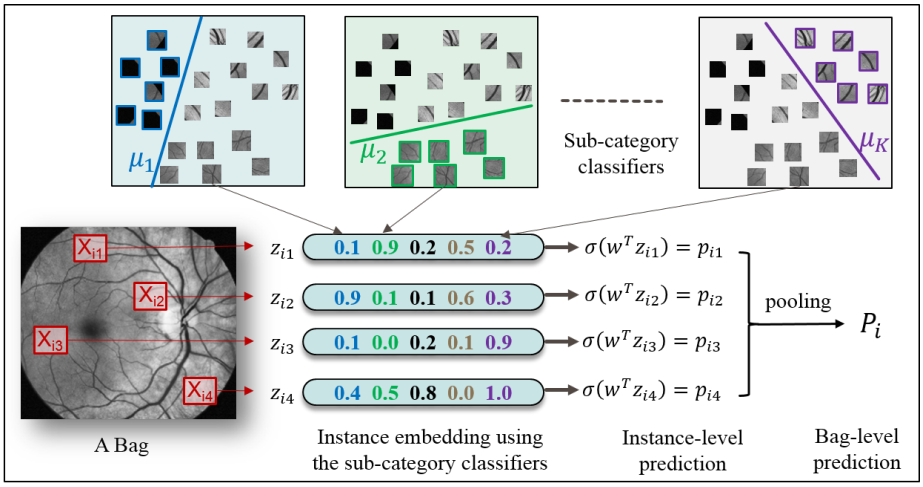


Fig. 2: Overview of the proposed approach.

3.2 Sub-category Classifiers for MIL

Let the training dataset contain $\{(B_i, y_i)\}_{i=1}^{N}$, where B_i is the i^{th} bag (image), $y_i \in \{-1, 1\}$ is its label, and N is the number of bags. Each bag B_i consists of N_i instances (image regions), so that $B_i = \{\mathbf{x}_{ij}\}_{j=1}^{N_i}$, where $\mathbf{x}_{ij} \in \mathbb{R}^d$ is the feature representation of the j^{th} instance of the i^{th} bag.

Let $\mathcal{M} = [\mu_1, \ldots, \mu_K] \in \mathbb{R}^{d \times K}$ be a set of sub-category classifiers, where each classifier is learned to separate a particular sub-category from others. The probability of an instance $\mathbf{x}_{ij}$ belonging to the k^{th} sub-category vs rest can be given as $q_{ijk} = \sigma(\mu_k^T \mathbf{x}_{ij})$, where $\sigma(\mathbf{x}) = 1/(1 + \exp(-\mathbf{x}))$. The new instance-representation $\mathbf{z}_{ij}$ in the discriminative sub-space is defined by the outputs of these sub-category classifiers, *i.e.* $\mathbf{z}_{ij} = [q_{ij1}, \ldots, q_{ijk}]$. Let $\mathbf{w} \in \mathbb{R}^K$ define the IL classifier which is learned in this discriminative subspace, and $p_{ij} = \sigma(\mathbf{w}^T \mathbf{z}_{ij})$ the probability of the instance $\mathbf{x}_{ij}$ belonging to the positive class.

The BL probability, P_i, of a bag B_i can be obtained by aggregating (pooling) the probabilities of the instances inside the bag. In this work, we use the *generalized-mean* operator ($\mathcal{G}$) for aggregation: $P_i = \left(\frac{1}{N_i}\sum_{j=1}^{N_i} p_{ij}^r\right)^{1/r}$, where r is a pooling parameter. When $r = 1$, $\mathcal{G}$ becomes *average-pooling*, and large r values ($r \to \infty$) approximate *max-pooling*.

The sub-category classifiers ($\mathcal{M}$), the pooling parameter (r), and the IL classifier ($\mathbf{w}$) can be learned using a cross-entropy loss function (Eq. (1)).

$$\underset{r,\mathcal{M},\mathbf{w}}{\arg\min} \ \frac{\lambda}{2}\|\mathbf{w}\|_2^2 - \frac{1}{N_+}\sum_{i:y_i=1}\log(P_i) - \frac{1}{N_-}\sum_{i:y_i=-1}\log(1-P_i) \qquad (1)$$

where $P_i = P_i(y_i = 1 | B_i, r, \mathcal{M})$, λ is a regularization parameter, and N_+, N_- are the total number of positive and negative bags in the training set respectively. Note that, this loss is widely used by the existing MIL approaches in [8,9,11,12].

Instead, we propose a margin-based loss (Eq. (2)) which penalizes the bags violating the margin, as margin-based loss has two main advantages over the cross-entropy loss [13]. (1) It tries to improve the classification accuracy of the training data (by focussing on the wrongly classified images), instead of making the correct predictions more accurate (as in cross-entropy loss). (2) It improves training speed, as model updates are only based on the images classified wrongly; the ones classified correctly will not contribute to the model updates, and can be avoided altogether in derivative calculations.

$$\begin{aligned} \arg\min_{r,\mathcal{M},\mathbf{w}} \ & \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{1}{N_+}\sum_{i:y_i=1}\mathcal{L}_i(y_i,B_i,r,\mathcal{M}) + \frac{1}{N_-}\sum_{i:y_i=-1}\mathcal{L}_i(y_i,B_i,r,\mathcal{M})) \\ & \text{where,} \quad \mathcal{L}_i(y_i,B_i,r,\mathcal{M}) = \max\left[0,\ \gamma + y_i(0.5-P_i)\right]^2. \end{aligned} \tag{2}$$

$\gamma \in (0, 0.5]$ is a margin parameter. In our experiments we set $\gamma = 0.1$, $\lambda = 10^2$.
Initialization and optimization: We use gradient descent to optimize Eq. (2), alternating between optimizing $\mathcal{M}$, $\mathbf{w}$ and r until convergence. To initialize $\mathcal{M}$, first the instances from the training set are clustered using k-means (dictionary size $= K$), then a set of one-vs-rest linear SVM classifiers are learned to separate each cluster from the rest. These classifiers give the initial values to $\mathcal{M}$.

4 Experiments

4.1 Datasets and Experimental Settings

The experimental settings for different datasets are summarized in Table 1.
(1) Messidor [3]: A public diabetic retinopathy screening dataset, contains 1200 eye fundus images. Well studied in [3] for BL classification. Each image was rescaled to 700×700 pixels and split into 135×135 regions. Each region was represented by a set of features including intensity histograms and texture.
(2) The diabetic retinopathy (DR) screening dataset [4]: 425 FC images, constructed from 4 publicly available datasets (DiabRetDB0, DiabRetDB1, STARE and Messidor). Each image is represented by a set of 48 instances.
(3) UCSB breast cancer [5]: 58 TMA H&E stained breast images (26 malignant, 32 benign). Used in [5, 8, 12] to compare different MIL approaches; each image was divided into 49 instances, and each instance is represented by a 708-dimensional feature vector including SIFT and local binary patterns.
(4) RNFL retinal fundus image dataset: Green channel was considered for processing. Images were resized preserving their aspect ratio so that their maximum dimension (row or column) becomes 700 pixels. Each image is then histogram-equalized. Instances (square image regions) of size 128×128 pixels with an overlap of 64 pixels are extracted, leading to ~ 90 instances per image. Inside each instance, SIFT features (patch size of 24×24 pixels, overlap 16 pixels) are computed and encoded using Sparse Coding with a dictionary size of 200. Average-pooling was applied to get a feature representation for each instance.

Dataset		no of images			Exp. setup
		positive	negative	total	
Messidor [3]		654	546	1200	10 times 2-FCV
DR [4]		265	160	425	fixed train($\frac{2}{3}$)-test($\frac{1}{3}$) split
UCSB Breast cancer [5]		26	32	58	10 times 4-FCV
RNFL fundus images	A1	348	228	576	20 times 2-FCV
	A2	436	140	576	

Table 1: Datasets and experimental settings (FCV-fold cross validation).

Method	Acc.
MI-SVM [6]	54.5
SIL-SVM	58.4
GP-MIL	59.2
MILBoost	64.1
mi-Graph [14]	72.5
Ours	**73.1**

(a) Messidor dataset [3]

Method	Acc.
mi-SVM [6]	70.32
MILES	71.00
EMDD	73.50
SNPM [4]	81.30
mi-Graph [14]	83.87
Ours	**88.00**

(b) DR dataset [4]

Method	AUC
MILBoost	0.83
MI-SVM [6]	0.90
BRT [12]	0.93
mi-Graph [14]	0.946
JC^{2}MIL [8]	0.95
Ours	**0.965**

(c) UCSB cancer [5]

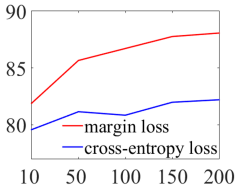
Table 2: Results on the public datasets. All the results except ours and mi-Graph were copied from [3–5]. Some references are omitted due to space. Different evaluation measures were used as they were reported in [3–5].

4.2 Experiments with Public Medical Image Datasets

Table 2 reports the comparative results on the public datasets. For fair comparison we use directly the features made publicly available [4], and follow the same experimental set-up used by the existing approaches.

With Messidor, our approach gives a competitive accuracy of 73.1% (with a standard error of ± 0.12) compared to the accuracy obtained by mi-Graph,which however cannot provide IL predictions as a BL approach. With DR, our approach improves the state-of-the-art accuracy by $\sim$ 4%. With UCSB, our approach achieves an AUC of 0.965 with a standard error of 0.001. Our Equal Error Rate was 0.07 ± 0.002, much smaller than the one reported in [12] (0.16 ± 0.03).

The figure on the right shows K (x-axis) vs. accuracy values (y-axis) for the DR dataset. As expected, increasing K improves the accuracy, saturating for $K > 150$. This figure also shows that the margin-based loss (Eq. (2)) outperforms the cross-entropy loss (Eq. (1)). The advantages of the margin-based loss are discussed in Section 3.



4.3 RNFL Visibility Classification

We used the public code from [3] for MILBoost and mi-SVM, taking care to select the parameters guaranteeing the fairest possible comparison. As a

[4] Messidor and UCSB cancer: `http://www.miproblems.org/datasets/`;
DR: `https://github.com/ragavvenkatesan/np-mil`

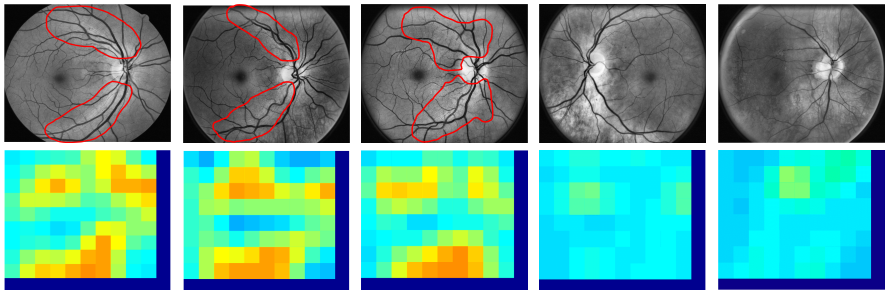


Fig. 3: Example region-level predictions for test images. Top row: Images with rough annotations for visible RNFL regions. In the last two images RNFL is invisible. Second row: Region-level probabilities obtained by the proposed approach, where the high values (red) indicate the probable RNFL visible regions.

Method	Percentage of images agreed ($\mathcal{P}$)		Kappa values ($\mathcal{K}$)	
	A1	A2	A1	A2
mi-SVM	73.11 ± 0.27	71.92 ± 0.62	0.4354 ± 0.0042	0.3942 ± 0.0063
MIL-Boost	80.53 ± 0.09	78.35 ± 0.09	0.5865 ± 0.0020	0.4940 ± 0.0029
BL-SVM	83.64 ± 0.09	80.72 ± 0.09	0.6523 ± 0.0017	0.5526 ± 0.0024
Ours	$\mathbf{83.78 \pm 0.08}$	$\mathbf{82.09 \pm 0.09}$	$\mathbf{0.6539 \pm 0.0017}$	$\mathbf{0.5798 \pm 0.0020}$

Table 3: Approaches and their agreements ($\mathcal{P}$ and $\mathcal{K}$ $\pm$ standard error) with different annotators (A1 and A2) for RNFL visibility classification. Note that the agreement between the two annotators is $\mathcal{P} = 82.99\%$ and $\mathcal{K} = 0.6190$.

further basline, we implemented BL-SVM, a supervised linear SVM classifier trained on the image-level feature representations obtained by average-pooling the dictionary-encoded (size 200) SIFT features. The training images with consensus labels from the annotators were used for training for each cross-validation.

Table 3 reports the results. Our approach gives better agreement with the annotators than other approaches. Over the entire RNFL dataset we found that the inter-annotator agreement is $\mathcal{P} = 82.99\%$ with a kappa value of $\mathcal{K} = 0.6190$. Our approach gives higher agreement with the experienced annotator (A1) than the less-experienced one (A2). Notice that, although BL-SVM gives a competitive performance compared to our approach, it cannot give region-level predictions as a BL method. Fig. 3 shows some region-level predictions by our approach.

5 Conclusions

The RNFL thickenss and its visibility have been posited as biomarkers for neurodegenerative conditions. We have proposed a novel MIL method to assess the visibility (visible/not visible) of the RNFL in fundus camera images, which would enable screening of much larger patient volumes than OCT. In addition, our approach locates visible RNFL regions from image-level training labels.

Experiments suggest that our margin-based loss solution performs better than the cross-entropy loss used by existing EB MIL approaches [8, 9, 12]. Ex-

periments with a local RNFL and 3 public medical image datasets show considerable improvements compared to the state-of-the-art. Future work will address the associations of RNFL visibility with brain features and patient outcome.
Acknowledgement: S. Manivannan is supported by EPSRC grant EP/M005976/1. The authors would like to thank Prof. Stephen J. McKenna and Dr. Jianguo Zhang for valuable comments.

References

1. Thomson, K.L., Yeo, J.M., Waddell, B., Cameron, J.R., Pal, S.: A systematic review and meta-analysis of retinal nerve fiber layer change in dementia, using optical coherence tomography. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring **1**(2) (2015) 136 – 143
2. Hedges, H.R., Galves, R.P., Speigelman, D., Barbas, N.R., Peli, E., Yardley, C.J.: Retinal nerve fiber layer abnormalities in Alzheimer's disease. Acta Ophthalmologica Scandinavica **74**(3) (1996) 271 – 275
3. Kandemir, M., Hamprecht, F.A.: Computer-aided diagnosis from weak supervision: A benchmarking study. Computerized Medical Imaging and Graphics **42** (2015) 44–50
4. Venkatesan, R., Chandakkar, P., Li, B.: Simpler non-parametric methods provide as good or better results to multiple-instance learning. In: IEEE International Conference on Computer Vision. (2015) 2605–2613
5. Kandemir, M., Zhang, C., Hamprecht, F.A.: Empowering Multiple Instance Histopathology Cancer Diagnosis by Cell Graphs. In: MICCAI. Part II, LNCS, Springer (2014) 228–235
6. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems 15, MIT Press (2003) 561–568
7. Xu, Y., Zhu, J.Y., Chang, E., Tu, Z.: Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition. (2012) 964–971
8. Sikka, K., Giri, R., Bartlett, M.: Joint clustering and classification for multiple instance learning. In: British Machine Vision Conference. (2015) 71.1–71.12
9. Ruiz, A., Van de Weijer, J., Binefa, X.: Regularized multi-concept MIL for weakly-supervised facial behavior categorization. In: British Machine Vision Conference. (2014)
10. Chen, Y., Bi, J., Wang, J.: MILES: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12) (2006) 1931–1947
11. Kraus, O.Z., Ba, L.J., Frey, B.J.: Classifying and segmenting microscopy images using convolutional multiple instance learning. CoRR **abs/1511.05286** (2015)
12. Li, W., Zhang, J., McKenna, S.J.: Multiple instance cancer detection by boosting regularised trees. In: MICCAI, Part I, LNCS, Springer (2015) 645–652
13. Jin, J., Fu, K., Zhang, C.: Traffic sign recognition with hinge loss trained convolutional neural networks. IEEE Transactions on Intelligent Transportation Systems **15**(5) (2014) 1991–2000
14. Zhou, Z.H., Sun, Y.Y., Li, Y.F.: Multi-instance learning by treating instances as non-i.i.d. samples. In: International Conference on Machine Learning. (2009) 1249–1256