Effective Drusen Localization for Early AMD Screening using Sparse Multiple Instance Learning

Huiying Liu, Yanwu Xu, Damon W.K. Wong, and Jiang Liu

Institute for Infocomm Research, A*STAR, Singapore

Abstract. Age-related Macular Degeneration (AMD) is one of the leading causes of blindness. Automatic screening of AMD has attracted much research effort in recent years because it brings benefits to both patients and ophthalmologists. Drusen is an important clinical indicator for AMD in its early stage. Accurately detecting and localizing drusen are important for AMD detection and grading. In this paper, we propose an effective approach to localize drusen in fundus images. This approach trains a drusen classifier from a weakly labeled dataset, i.e., only the existence of drusen is known but not the exact locations or boundaries, by employing Multiple Instance Learning (MIL). Specifically, considering the sparsity of drusen in fundus images, we employ sparse Multiple Instance Learning to obtain better performance compared with classical MIL. Experiments on 350 fundus images with 96 having AMD demonstrates that on the task of AMD detection, multiple instance learning, both classical and sparse versions, achieve comparable performance compared with fully supervised SVM. On the task of drusen localization, sparse MIL outperforms MIL significantly.

1 Introduction

Age-related Macular Degeneration (AMD), after cataracts and glaucoma, is the third leading cause of blindness worldwide and the principal cause in the elderly [1]. Population based AMD screening is a necessity because of its increasing prevalence. However, symptoms (e.g., vision scotoma and distortion) in the early stage of AMD are generally not obvious to those who have them, and thus patients usually find out late that they contracted AMD. As a result, the vision of those patients may be permanently lost. Therefore, it is of great importance to detect AMD at the early stage, so that patients can undergo treatment to prevent it from worsening. For patients already diagnosed with AMD, accurate grading is important to customize a suitable treatment. Regular screening is a potential way to detect and grade early AMD. However, manual detection and grading of AMD are time consuming and labor intensive. More importantly, because grading is usually subjective (graded by human clinicians), a certain degree of inaccuracy may be introduced.

Automatic AMD screening brings benefits to both patients and ophthalmologists, hence it has attracted much research effort. The main sign of early AMD is the appearance of drusen [2]. Drusen appears as a yellow-white spot in digital fundus imagery, which is the most popular imaging modality for AMD diagnosis. Examples of drusen are shown in Figure 1 (b). Drusen detection has significant importance for early AMD detection and grading.

Because of the importance of automatic AMD screening, many methods have been proposed for AMD detection in recent years. They can be separated into roughly two
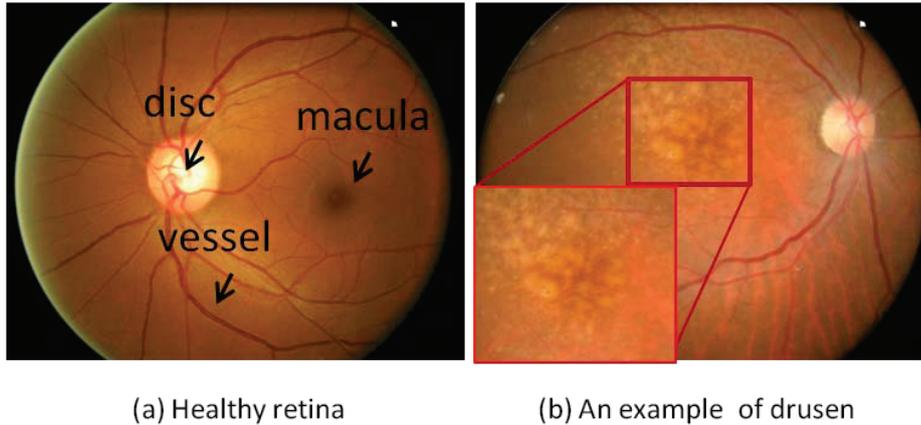
(a) Healthy retina   (b) An example of drusen

Fig. 1. Examples of fundus images with and without drusen.

classes, bottom-up methods and top-down methods. Bottom-up methods detect AMD though drusen segmentation, and can be further classified into three major categories. The first category, consisting of the earliest drusen segmentation methods, is based on local maxima, e.g., the geodesic method [3]. These methods first detect local maxima, then further classify the candidates according to contrast, size and shape. The second category consists of local threshold based methods, e.g., Histogram based Adaptive Local Thresholding (HALT) [4], and Otsu based adaptive thresholding [5]. The third category includes methods that operate in the frequency domain, e.g., wavelets [6], and that involve amplitude-modulation frequency modulation (AM-FM) [7].

These bottom-up AMD detection methods are restricted by the accuracy of drusen segmentation. To bypass drusen segmentation, in recent years, researchers have started to seek methods that detect AMD directly from fundus images, without drusen segmentation. These methods describe an image using locally extracted features, then the features are fed into a classifier to decide whether the image contains drusen or not. An early attempt in this direction used a histogram based representation followed by Case-Based Reasoning [8]. The latest methods in this category include the Biologically Inspired Feature (BIF) based method [9] and the Hierarchical Word Image representation and SPIN features based method called Thalia [10].

Although the methods for detecting AMD without drusen segmentation have shown some progress, drusen localization/segmentation has its importance in AMD grading, which is even more important than AMD detection. The severity of AMD is usually measured according to position, size, and type (hard or soft), together with geographic atrophy and other symptoms. The overlap of drusen with macula is used to measure the severity of AMD [11]. Machine learning is a feasible way to improve the performance of drusen localization/segmentation but it needs labor-intensive manually marked data as training data. To reduce the need for training data, we propose to utilize multiple instance learning to train a drusen classifier from weakly labeled data, meaning that each image is labeled as having drusen or not but without a known location or boundary of the drusen. We observed that for drusen classification, the positive bags are mostly

sparse, meaning that drusen compose only a small portion of the detected candidates. We employ sparse multiple instance learning to fit the sparseness of this problem.

The contribution of this paper is that we propose an effective drusen localization approach for AMD screening. The two major characteristics of this approach are as follows. 1) It is able not only to detect AMD but also to track the locations of drusen for further AMD grading. 2) It employs Multiple Instance Learning (MIL) to train the detectors from a weakly labeled dataset, thus significantly reducing the need for manual labeling. Specifically, it employs sparse MIL to fit the sparse distribution of drusen.

2   Drusen localization using sparse MIL

For the task of drusen localization, we follow the steps of 1) drusen candidate detection, 2) feature extraction, and 3) drusen classification using the learned classifier. We first state the problem formulation of training the classifier using sparse multiple instance learning.

2.1   Problem formulation

We formulate drusen localization as a MIL problem. Specifically, each fundus image is a bag, each detected drusen candidate is an instance. Only the bag labels are available, which indicate the existence of drusen or not. A bag is labeled as positive if at least one sample in it is positive (drusen). Conversely it is labeled as negative if all the samples in it are negative (not drusen). The instance labels are not available for training.

To transform the weakly supervised learning problem to a supervised one, a bag is represented as the sum of all its instances, normalized by its 1 or 2-norm. This is the Normalization Set Kernel (NSK). If we consider the fact that all the instances in the negative bags are negative, we obtain the tight form of the NSK optimization problem:

$$
\begin{aligned}
\min_{\mathbf{w},b,\xi_{\mathbf{x}},\xi_{\mathbf{X}}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{|\tilde{\mathcal{X}}_n|}\sum_{\mathbf{x}\in\tilde{\mathcal{X}}_n}\xi_{\mathbf{x}} + \frac{C}{|\mathcal{X}_p|}\sum_{\mathbf{X}\in\mathcal{X}_p}\xi_{\mathbf{X}} \\
\text{s.t.} \quad & \mathbf{w}^T\phi(\mathbf{x})+b \leq -1+\xi_{\mathbf{x}}, \forall \mathbf{x}\in\tilde{\mathcal{X}}_n, \\
& \mathbf{w}^T\frac{\phi(X)}{|X|}+b \geq 1-\xi_X, \forall X\in\mathcal{X}_p, \\
& \xi_{\mathbf{x}}\geq 0,\ \xi_X\geq 0.
\end{aligned}
\tag{1}
$$

Here χ is the set of training bags, $\chi_p\subseteq\chi$ and $\chi_n\subseteq\chi$ are the sets of positive bags and negative bags, respectively. $\tilde{\chi}_p=\{\mathbf{x}|\mathbf{x}\in X\in\chi_p\}$ and $\tilde{\chi}_n=\{\mathbf{x}|\mathbf{x}\in X\in\chi_n\}$ are the sets of instances from positive bags and negative bags, respectively. Let $\phi(\mathbf{x})$ be the feature vector representation of $\mathbf{x}$. $\phi(X)=\sum_{\mathbf{x}\in X}\phi(\mathbf{x})$ is the feature vector representation of bag X. $w\phi(\mathbf{x})+b$ is the SVM decision hyperplane, parameterized by a weight vector w and bias b.

For every positive bag, the learned decision function is used to select the bag instance that gives the maximum value, and the bag representation is replaced with this instance. The SVM classifier is then retrained with the new dataset. This process is repeated until no bag representation is changed. This is the MI-SVM algorithm.

The inequality constraint on positive bags in NSK can be seen as a balancing constraint. Let $y(\mathbf{x}) = \pm 1$ be the hidden label of an instance $\mathbf{x}$ from a positive bag X. Then the inequality constraint in NSK can be rewritten as follows:

$$\sum_{\mathbf{x} \in X} \frac{w\phi(\mathbf{x}) + b}{|X|} \geq \sum_{\mathbf{x} \in X} \frac{y(\mathbf{x})}{|X|} - \xi_{\mathbf{x}}$$

$$y(\mathbf{x}) = 1, \forall \mathbf{x} \in X.$$

$$\text{(2)}$$

The balancing constraint above is not consistent with the real data distribution, since it implicitly assumes that all instances inside the bag are positive. This is especially problematic when the bag X is sparse in positive instances. In our situation, drusen classification, the detected candidates are usually very noisy because many points near blood vessels are also detected as maximum points. For example, in the ACHIKO-D350 dataset [12], the average drusen rate in the 96 fundus images with drusen is only 8.95%. In 53% of the images, positive samples occupy less than 5% of the detected candidates. Fig. 2 shows an example of the detected candidates. Therefore, we would like the constraint to express the requirement that at least one instance $\hat{\mathbf{x}}$ from the bag is positive, as shown below:

$$\sum_{\mathbf{x} \in X} \frac{w\phi(\mathbf{x}) + b}{|X|} \geq \sum_{\mathbf{x} \in X} \frac{y(\mathbf{x})}{|X|} - \xi_{\mathbf{x}}$$

$$y(\mathbf{x}) = -1, \forall \mathbf{x} \in X \setminus \{\hat{\mathbf{x}}\}$$

$$y(\hat{\mathbf{x}}) = +1.$$

$$\text{(3)}$$

Replacing the constraint function in Equation (1) with Equation (3) leads to the formulation of sparse Multiple Instance Learning (sMIL): [13].

$$\min_{w,b,\xi_{\mathbf{x}},\xi_{\mathbf{X}}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{|\tilde{\mathcal{X}}_n|} \sum_{\mathbf{x} \in \tilde{\mathcal{X}}_n} \xi_{\mathbf{x}} + \frac{C}{|\mathcal{X}_p|} \sum_{\mathbf{X} \in \mathcal{X}_p} \xi_{\mathbf{X}}$$

$$\text{s.t.} \quad \mathbf{w}^T \phi(\mathbf{x}) + b \leq -1 + \xi_{\mathbf{x}}, \forall \mathbf{x} \in \tilde{\mathcal{X}}_n,$$

$$\mathbf{w}^T \frac{\phi(\mathbf{X})}{|\mathbf{X}|} + b \geq \frac{2 - |X|}{|X|} - \xi_X, \forall X \in \mathcal{X}_p,$$

$$\xi_{\mathbf{x}} \geq 0, \ \xi_{\mathbf{X}} \geq 0.$$

$$\text{(4)}$$

The item $\frac{2-|X|}{|X|}$ adds a weight to each bag. In this setting, a smaller bag will have a larger weight under the assumption that all the bags contain a similar number of positive samples. This setting is especially useful when the positive bags are sparse.

2.2   Bag construction

Drusen candidate detection Drusen appear as white or yellow spots in fundus images, so drusen are local maximum points. To construct the bags for MIL, we first detect local maximum points as drusen candidates within the Gaussian space, which is constructed by convolving the image with a Gaussian function at each level:

$$L(i, j, \sigma_k) = G(i, j, \sigma_k) * I(i, j),$$
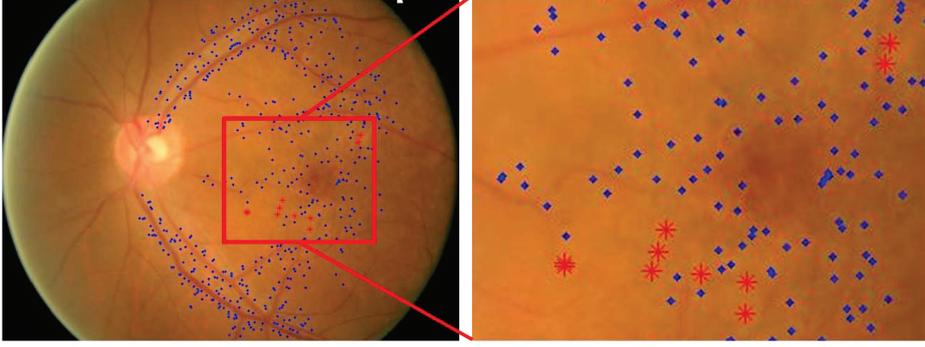
$$\text{(5)}$$

Fig. 2. An example of sparse bag. True and false drusen points are marked with red stars and blue dots, respectively. The drusen points compose 2% of the candidates.

where I is the input image, (i, j) is the position in the image, $G(i, j, \sigma_k)$ is a Gaussian function with standard deviation of σ_k, and $*$ is the convolution operator. Then from the Gaussian space, we construct the Laplacian space as follows:

$$
\begin{aligned}
D(i, j, \sigma_k) &= (G(i, j, \sigma_k) - G(i, j, \sigma_{k+1})) * I(i, j) \\
&= L(i, j, \sigma_k) - L(i, j, \sigma_{k+1}).
\end{aligned}
\tag{6}
$$

The Laplacian space has one fewer level than the Gaussian space. We adopt $\sigma = \sqrt{2}^{\{1,2,\dots,8\}}$. A pixel is determined to be a maximum if it is a local maximum in the Laplacian space. The local window used to determine the maximum is $3 \times 3 \times 3$. Thus, we detect maximum points at scales $\sigma = \sqrt{2}^{\{2,3,\dots,6\}}$. In Equation (6), $D(i, j, \sigma) = 0$ when $i^2 + j^2 = \frac{2a^2\sigma^2 \log(a^2)}{a^2-1}$, where $a = \frac{\sigma_{k+1}}{\sigma_k}$. When $a = \sqrt{2}$, the radius of detected drusen at scale σ is $r = 1.67\sigma$. Thus the minimal and maximal diameters of detected drusen are 5 and 27 pixels respectively. This process will detect about 4000 maxima points in a fundus image of size 1024x683. The detected points are first selected according to intensity, contrast, and distance to macula, to reduce data size.

Feature description The characteristics distinguishing drusen from the background are intensity, shape, size, contrast and texture. Further, drusen usually distribute around macula. The features adopted to describe these characteristics are 1) edge direction histogram, 2) intensity, 3) contrast, 4) size, 5) Hessian value, 6) distance to macula, 7) Gabor coefficients, and 8) SPIN feature [14].

Drusen classification and AMD detection After constructing the bags and extracting the features, we use the trained model to classify each drusen candidate (instance $\mathbf{x}$) as drusen or not. With the classification result of instance $\mathbf{x}$ denoted as $y(\mathbf{x})$, the probability of an image (bag X) is calculated as $y(X) = \frac{1}{|X|} \sum_{\mathbf{x} \in X, y(\mathbf{x})=1} y(\mathbf{x})$.

3 Experiments

3.1 Dataset

The experiments are performed on ACHIKO-D350 [12], which is, as far as we know, the only dataset with drusen boundaries marked. It consists of 350 population-based

Table 1. Result of comparison between SVM, MIL, and sparse MIL.

	SVM		MI-SVM		sMIL	
	mean	std	mean	std	mean	std
AUC of AMD detection (bag level)	0.97	0.01	0.95	0.03	0.96	0.03
AP of AMD detection (bag level)	0.93	0.01	0.93	0.06	0.93	0.05
AP of drusen classification (instance level)	0.79	0.02	0.69	0.02	0.72	0.03

images from the Singapore Eye Malay Study, consisting of 96 clinically verified drusen images and 254 non-drusen images. Each image has a resolution of 3072x2048. Besides drusen boundaries, the position of the macula is also manually marked. In experiments, the images are resized to 1024x683.

3.2   Measurements

The measurements adopted in the experiments include Average Precision and the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. Average Precision is the average of sensitivity and specificity. ROC curve is the true positive rate (sensitivity) vs. false positive rate (1-specificity) curve. Sensitivity and specificity are defined in Eq. (7).

$$Sensitivity = \frac{True\,positives}{Positives}, Specificity = \frac{True\,negatives}{Negatives}. \qquad (7)$$

3.3   Comparison with baseline

We compare the performance of using sparse MIL with the baseline method, meaning using general MIL, i.e., MI-SVM. Furthermore, we compare the method with SVM, which uses additional drusen labels. SVM is regarded as an ideal reference of the best performance that can be achieved by learning from images without drusen labels, and is used to validate the effectiveness of the proposed approach using sMIL. We adopt libSVM [15] to implement these algorithms. All the three methods use the same kernel type, specifically the Radial Basis Function (RBF) kernel with cost parameter set to 1. Comparison results are shown in Table 1. Fig. 3 shows two examples of drusen classi-fication results. The top row shows a failure case in which MIL and sMIL obtain much worse performance than SVM does. In this case, the drusen are clumped together and thus are classified as negative (non-drusen) by MIL and sMIL. The bottom row shows a successful case in which MIL and sMIL obtain performance comparable to SVM.

We should note that since MIL and sMIL take SVM as a basis and use less labeling information than SVM does, the performance of SVM is the theoretical upper bound of MIL and sMIL. From the result, we can draw the following conclusions.

1. On the task of AMD detection, MIL and sMIL achieve performance comparable to SVM in terms of AUC and average precision. We can also see that sMIL outper-forms MIL slightly.
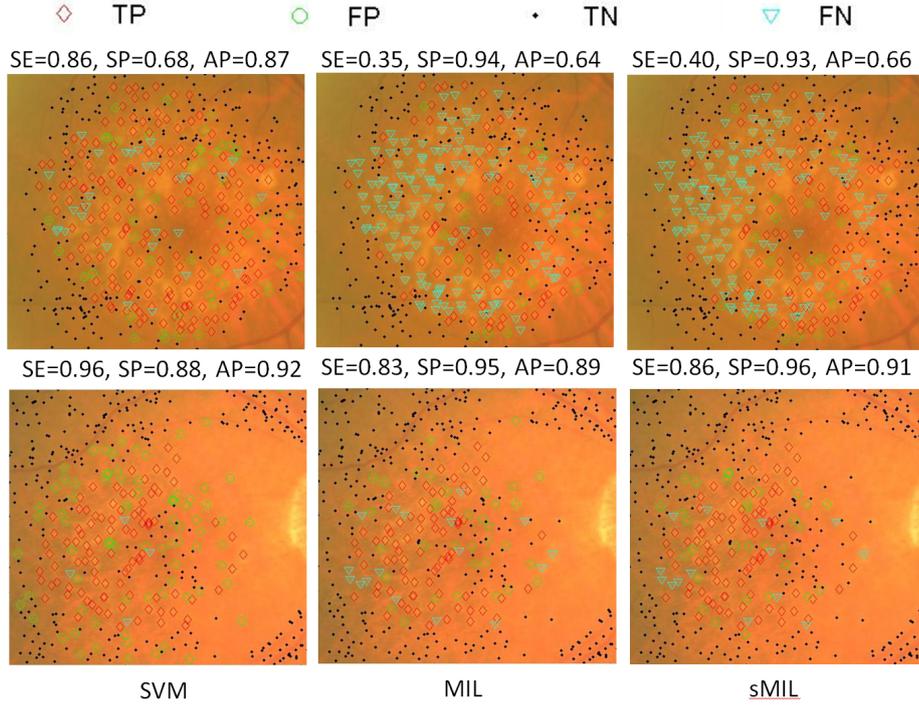
◇ TP   ◯ FP   · TN   ▽ FN

| SE=0.86, SP=0.68, AP=0.87 | SE=0.35, SP=0.94, AP=0.64 | SE=0.40, SP=0.93, AP=0.66 |

| SE=0.96, SP=0.88, AP=0.92 | SE=0.83, SP=0.95, AP=0.89 | SE=0.86, SP=0.96, AP=0.91 |

| SVM | MIL | sMIL |

Fig. 3. Examples of drusen classification results. Top row shows a failure case in which MIL and sMIL obtain lower sensitivity. Bottom row shows a successful case where MIL and sMIL obtain results comparable to SVM.

2. On the task of drusen classification, sMIL outperforms MIL significantly (with a significance test result of $p < 0.01$).

The above results verify that 1) MIL is an effective method for drusen localization and 2) sparse MIL is more suitable for this task than general MIL.

3.4 Comparison with State-of-the-Art

As reported in their papers, the BIF based method [9] and the Thalia system [10] achieved average precision of 0.89±0.01 and 0.95±0.01, respectively, on the same dataset. Our method achieves performance comparable to these two methods on the task of AMD detection. These two methods can only perform AMD detection but not drusen localization.

4 Conclusion

In this paper, we propose an effective drusen localization method for AMD detection. This approach employs multiple instance learning to reduce the need for manual labeling of training data. Specifically, it employs sparse multiple instance learning to fit the

properties of the drusen classification problem. The proposed method is experimentally verified to be effective, especially in the task of drusen classification. One important application of drusen localization and segmentation is AMD grading. In our future work, we will extend our approach to grade the severity of AMD.

References

1. Kawasaki, R., Yasuda, M., Song, S.J., Chen, S.J., Jonas, J.B., Wang, J.J., Mitchell, P., Wong, T.Y.: The prevalence of age-related macular degeneration in asians: a systematic review and meta-analysis. Ophthalmology **117**(5) (2010) 921–927
2. De Jong, P.T.: Age-related macular degeneration. New England Journal of Medicine **355**(14) (2006) 1474–1485
3. Ben Sbeh, Z., Cohen, L.D., Mimoun, G., Coscas, G.: A new approach of geodesic reconstruction for drusen segmentation in eye fundus images. IEEE Transactions on Medical Imaging **20**(12) (2001) 1321–1333
4. Rapantzikos, K., Zervakis, M., Balas, K.: Detection and segmentation of drusen deposits on human retina: Potential in the diagnosis of age-related macular degeneration. Medical image analysis **7**(1) (2003) 95–108
5. Smith, R.T., Chan, J.K., Nagasaki, T., Ahmad, U.F., Barbazetto, I., Sparrow, J., Figueroa, M., Merriam, J.: Automated detection of macular drusen using geometric background leveling and threshold selection. Archives of ophthalmology **123**(2) (2005) 200
6. Brandon, L., Hoover, A.: Drusen detection in a retinal image using multi-level analysis. In: MICCAI. Springer (2003) 618–625
7. Barriga, E., Murray, V., Agurto, C., Pattichis, M., Russell, S., Abramoff, M., Davis, H., Soliz, P.: Multi-scale am-fm for lesion phenotyping on age-related macular degeneration. In: 22nd IEEE International Symposium on Computer-Based Medical Systems, IEEE (2009) 1–5
8. Hijazi, M.H.A., Coenen, F., Zheng, Y.: Retinal image classification using a histogram based approach. In: IEEE International Joint Conference on Neural Networks. (2010) 3501–3507
9. Cheng, J., Wong, D.W.K., Cheng, X., Liu, J., Tan, N.M., Bhargava, M., Cheung, C.M.G., Wong, T.Y.: Early age-related macular degeneration detection by focal biologically inspired feature. In: ICIP, IEEE (2012) 2805–2808
10. Wong, D.W., Liu, J., Cheng, X., Zhang, J., Yin, F., Bhargava, M., Cheung, G.C., Wong, T.Y.: Thalia-an automatic hierarchical analysis system to detect drusen lesion images for amd assessment. In: ISBI, IEEE (2013) 884–887
11. Medhi, J.P., Nath, M.K., Dandapat, S.: Automatic grading of macular degeneration from color fundus images. In: Information and Communication Technologies (WICT), 2012 World Congress on, IEEE (2012) 511–514
12. Liu, H., Xu, Y., Wong, D.W.K., Laude, A., Lim, T.H., Liu, J.: Achiko-d350: A dataset for early amd detection and drusen segmentation. In: Ophthalmic Medical Image Analysis (MICCAI workshop). (2014)
13. Bunescu, R.C., Mooney, R.J.: Multiple instance learning for sparse positive bags. In: Proceedings of the 24th international conference on Machine learning, ACM (2007) 105–112
14. Johnson, A.E., Hebert, M.: Using spin images for efficient object recognition in cluttered 3d scenes. IEEE Trans. on Pattern Analysis and Machine Intelligence, **21**(5) (1999) 433–449
15. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27 Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.