Automated Foveola Localization in Retinal 3D-OCT Images Using Structural Support Vector Machine Prediction

Yu-Ying Liu[1], Hiroshi Ishikawa[2,3], Mei Chen[4], Gadi Wollstein[2],
Joel S. Schuman[2,3], and James M. Rehg[1]

[1] College of Computing, Georgia Institute of Technology, Atlanta, GA
[2] UPMC Eye Center, University of Pittsburgh School of Medicine, Pittsburgh, PA
[3] Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA
[4] Intel Science and Technology Center on Embedded Computing, Pittsburgh, PA

Abstract. We develop an automated method to determine the foveola location in macular 3D-OCT images in either healthy or pathological conditions. Structural Support Vector Machine (S-SVM) is trained to directly predict the location of the foveola, such that the score at the ground truth position is higher than that at any other position by a margin scaling with the associated localization loss. This S-SVM formulation directly minimizes the empirical risk of localization error, and makes efficient use of all available training data. It deals with the localization problem in a more principled way compared to the conventional binary classifier learning that uses zero-one loss and random sampling of negative examples. A total of 170 scans were collected for the experiment. Our method localized 95.1% of testing scans within the anatomical area of the foveola. Our experimental results show that the proposed method can effectively identify the location of the foveola, facilitating diagnosis around this important landmark.

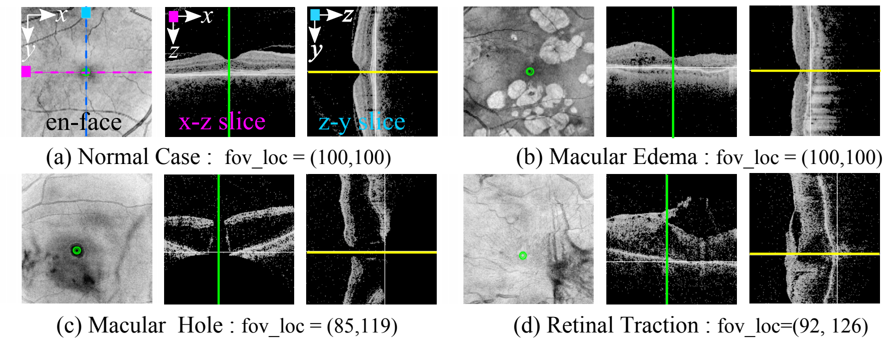
1 Introduction

The foveola is an important anatomical landmark for retinal image analysis [1]. It is located in the center of the macula, responsible for sharp central vision. Several clinically-relevant indices are measured with respect to the foveola location, such as the retina's average thickness, or drusen size within concentric circles around the foveola [1, 2]. In addition, many macular diseases are best observed around the foveola, such as macular hole, and age-related macular degeneration [3]. Therefore, the localization of the foveola in retinal images is an important first step for diagnosis and longitudinal data analysis.

There has been extensive work in determining the foveola location in 2D color fundus images [1, 4]. However, there has been *no published work on automated foveola localization in retinal 3D-OCT images*. Researchers in ophthalmology typically need to determine this landmark in 3D-OCT images manually [2, 3].

Examples of the foveola location in 3D-OCT images are illustrated in Fig. 1, where the *OCT en-face* is a 2D image generated by projecting the 3D-OCT volume along the z (depth) axis, a x-y plane analogous to the *fundus* image.

N. Ayache et al. (Eds.): MICCAI 2012, Part I, LNCS 7510, pp. 307–314, 2012.



(a) Normal Case : fov_loc = (100,100)

(b) Macular Edema : fov_loc = (100,100)

(c) Macular Hole : fov_loc = (85,119)

(d) Retinal Traction : fov_loc=(92, 126)

Fig. 1. Examples of the foveola's (x,y) location in normal and diseased cases. On the en-face image, the (x, y) location is marked by a *green* circle, while in the corresponding x-z (horizontal) and z-y (vertical) slice, the x and y location is shown in *green* and *yellow* line, respectively. (The 3D scan is normalized to 200x200x200 dimension.)

From Fig. 1, we can see that the localization task is not trivial, since the foveola can have significant appearance changes due to various ocular diseases.

In the literature, an object localization or detection task is usually formulated as a discriminative binary classification problem [5–7], where *zero-one* loss term is employed. In training, the annotated ground truth locations form the positive set, while a number of negative examples are typically randomly-sampled. Each negative example is treated *equally negative*, regardless of its distance or area of overlap to the ground truth. Thus, the loss function used in training may not be the same as the one utilized in performance evaluation in testing (e.g. the Euclidean distance). This scheme has been applied to localizing organs in whole-body scans [5] and detection of liver tumors [6].

Recently Blaschko et. al [8] proposed to pose the object localization task as a *structural prediction* problem. Specifically, they adopted Structural Support Vector Machine (S-SVM) formulation [9, 10] to directly predict the coordinates of the target object's bounding box in a 2D image. S-SVM learns to predict outputs that can be a multivariate structure. The relationship between a possible output and the ground truth is explicitly modeled by a desired *loss* function. During training, the constraints state that the score at the ground truth should be higher than that of any other output by a required margin set to the loss term [9]. This formulation considers all possible output locations during training, and directly minimizes the empirical risk of localization. They have shown that the S-SVM outperforms binary classification for object localization in several 2D image datasets. However, *S-SVM has not yet been applied to medical image analysis.*

In the context of our task, the output space is the space of possible locations of the foveola in the 3D-OCT scan, which makes this problem a multivariate structural prediction problem. We adopt S-SVM framework to directly minimize the localization risk during training. A coarse-to-fine sliding window search approach is proposed to efficiently find the most-violated constraint in S-SVM's cutting-plane training and in prediction. In feature construction, multi-scale spatially-distributed texture features are designed to encode the appearance in the neighborhood of any candidate 3D position. We conducted experiments to compare S-SVM's performance with a human expert, and with the binary SVM classifier to validate our approach.

This paper makes three main contributions: (1) Introduce a formulation of the foveola localization problem in 3D-OCT as structured output prediction, which can be solved using S-SVM method. (2) Propose a coarse-to-fine sliding window-based approach to identify the most-violated constraint during S-SVM training. (3) Demonstrate high prediction accuracy using a dataset of 170 scans.

2 Approach

2.1 Formulation of Structural SVM in Foveola Localization Task

For our task, the optimization problem is formulated as follows: given a set of training scans $(a_1, ..., a_n) \subset A$ and the annotated foveola locations $(b_1, ..., b_n) \subset B$, the goal is to learn a function $g : A \mapsto B$ with which we can automatically label novel images. Note that since there is no consensus in defining the z (depth) location of the foveola in ophthalmology, we consider the output space B consisting of only the (x, y) labels. The extent of the retina in z direction can be estimated by a separate heuristic procedure and serves as an input for feature extraction (explained in Section 2.3).

The mapping g is learned by using the structured learning formula [9] as

$$g(a) = \text{argmax}_b \; f(a, b) = \text{argmax}_b \; \langle w, \phi(a, b) \rangle \tag{1}$$

where $f(a, b) = \langle w, \phi(a, b) \rangle$ is a linear discriminant function that should give a large score to pair (a, b) if they are well-matched, $\phi(a, b)$ is a feature vector associating input a and output b, and w is the weight vector to be learned. To learn w, we use the following *1-slack margin-rescaling* formulation of S-SVM [9],

$$\min_{w,\xi \geq 0} \; \frac{1}{2} w^T w + C\xi \tag{2}$$

$$s.t. \;\; \forall (\bar{b_1}, ..., \bar{b_n}) \in B^n : \frac{1}{n} \sum_{i=1}^{n} [\langle w, \phi(a_i, b_i) \rangle - \langle w, \phi(a_i, \bar{b_i}) \rangle] \geq \frac{1}{n} \sum_{i=1}^{n} \Delta(b_i, \bar{b_i}) - \xi \tag{3}$$

where $\Delta(b_i, \bar{b_i})$ is the loss function relating the two outputs, and is set to $\|b_i - \bar{b_i}\|_2$ representing their Euclidean distance, ξ is the slack variable, and C is a free parameter that controls the tradeoff between the slack and model complexity.

The constraints state that for each training pair (a_i, b_i), the score $\langle w, \phi(a_i, b_i) \rangle$ for the correct output b_i should be greater than the score of *all* other outputs $\bar{b_i}$ by a required margin $\Delta(b_i, \bar{b_i})$. If the margin is violated, the slack variable ξ becomes non-zero. In fact, ξ is the upper bound of the empirical risk on the training set [9], and is directly minimized in the objective function.

Algorithm 1. S-SVM training with margin-rescaling and 1-slack [9]

```
Input: Examples $S = \{(a_1, b_1), ..., (a_n, b_n)\}$, $C$, $\epsilon$;  Init: Constraints $W \leftarrow \emptyset$
Do
  $(w, \xi) \leftarrow \text{argmin}_{w,\xi \geq 0} \;\; \frac{1}{2} w^T w + C\xi$
   $s.t. \; \forall(\bar{b_1}, ..., \bar{b_n}) \in W : \frac{1}{n} \sum_{i=1}^{n} w^T [(\phi(a_i, b_i) - \phi(a_i, \bar{b_i})] \geq \frac{1}{n} \sum_{i=1}^{n} \Delta(b_i, \bar{b_i}) - \xi$
    For $i = 1, ..., n$
        $\bar{b_i} = \text{argmax}_b \;\; [w^T \phi(a_i, b) + \Delta(b_i, b)]$
    End for
    $W \leftarrow W \cup \{(\bar{b_1}, ..., \bar{b_n})\}$
Until $\frac{1}{n} \sum_{i=1}^{n} w^T [\phi(a_i, b_i) - \phi(a_i, \bar{b_i})] \geq \frac{1}{n} \sum_{i=1}^{n} \Delta(b_i, \bar{b_i}) - \xi - \epsilon$
Return  $(w, \xi)$
```

Note that the number of constraints in Eq. (3) is intractable, with the total number of constraints in $O(|B|^n)$. By using the *cutting-plane* training algorithm [9] (presented in Algo. 1 for completeness) that employs constraint-generation techniques, this large-scale optimization problem can be solved efficiently. Briefly, the weight vector w is estimated using a working set of constraints W which is set to *empty* initially, and new constraints are then added by finding the $\bar{b_i}$ for each a_i that violates the constraint the most (i.e., has the highest sum of the score function and the loss term). These two steps are alternated until no constraint can be found that is violated by more than the desired *precision* ϵ. This generally ends with a small set of active constraints [9]. Note that when the algorithm terminates, *all* constraints in B^n are satisfied within precision ϵ.

2.2 Finding the Most-Violated Constraint and Prediction

Note that in Algo. 1, we need an efficient method to find $\bar{b_i} = \text{argmax}_b \;\; [w^T \phi(a_i, b) + \Delta(b_i, b)]$ for each a_i, so as to construct the next constraint. Similarly, in prediction, it is desirable to efficiently derive $\hat{b} = \text{argmax}_b \; \langle w, \phi(a, b) \rangle$ for a novel input a. Previous work [8] addressed the above problems using a branch-and-bound procedure which exploited a bag-of-words feature model. Unfortunately such a technique cannot be easily adapted for the dense feature vectors (Section 2.3) needed for OCT image analysis. As an alternative, we propose to use a *coarse-to-fine sliding window search* approach to approximately obtain the desired result. Specifically, we first search the entire output range (x=[1 200], y=[1 200]) with 16-pixel spacing in both x and y, to identify the coarse position with the maximum score. The subsequent search ranges are $\pm 48, \pm 8, \pm 4$ in both x and y, with the sliding window centered around the previously found best location, at 4, 2, and 1 pixel spacing, respectively. A similar search strategy has been used for object detection [7] with a conventional classifier for improving the search speed.

2.3 Image Pre-processing and Feature Construction

We now describe the construction of our feature vector $\phi(a, b)$. First, before we can reliably extract features from a raw scan, a necessary pre-processing is

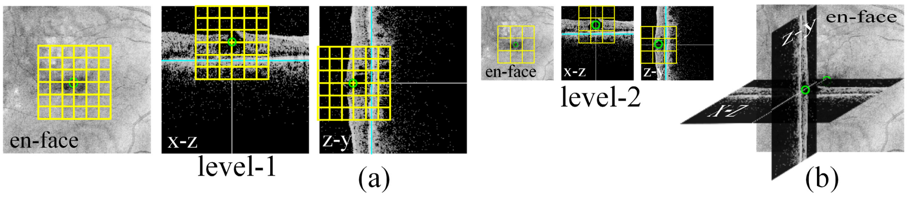


Fig. 2. (a) Illustration of our multi-scale spatially-distributed feature encoding for a given position (x, y, z). A 6x6 and 3x3 spatial grid is centered at the corresponding position on the en-face image, x-z slice and z-y slice for image scale level-1 and level-2, respectively. Appearance features are computed for each spatial cell. The automatically identified z position of the RPE layer is shown as a *light blue* line. (b) 3D presentation of the three orthogonal images (en-face, x-z slice, z-y slice) for a given 3D position.

to conduct eye-motion correction for restoring the 3D integrity of the volume. We apply Xu's [11] method to correct the eye motion artifacts, which usually produces a corrected volume with a roughly *flattened* retinal pigment epithelium (RPE) layer (the bottom retinal layer that shows high intensity in OCT images). This effect largely reduces the appearance variations across scans caused by different retinal curvatures or imagining deviation.

Before we can extract a volumetric feature vector centered at a candidate foveola location (x,y), we need to decide the retina's spatial extent in z. We now describe an empirical procedure to identify the maximum z value, $\hat{z}$, for analysis. We begin by estimating an average z position of the RPE layer in the volume. For each x-z slice, we find one row z that has the maximum average energy in the slice. This is usually located at the bottom RPE layer, but could sometimes map to the top nerve fiber layer. Then, the maximum z value among all x-z slices is found, and only the z within a specified distance to this maximum are retained, in order to exclude outliers. The z location of the RPE layer is estimated by taking the average of these retained z values. We found that this procedure can robustly derive the desired results (*light blue* line in Fig. 2(a)). We then set $\hat{z} = (z\_RPE + \frac{1}{10} dim\_z)$ as the largest z position for further analysis.

In constructing the feature $\phi(a, b)$ for a candidate output $b = (x, y)$, we compute features within the neighborhood centered at (x, y, z), where $z = (\hat{z} - \frac{1}{4} dim\_z)$. Specifically, we calculate features in the *three orthogonal context windows* (in en-face, x-z slice, and z-y slice) centered at (x, y, z). The window width/height is set to be $\frac{1}{2} dim\_size$ for each dimension. For each window, we divide it into 6x6 spatial cells, and compute intensity mean and gradient orientation histogram [12] with 16 angular bins for each cell. The same feature types are also computed for the down-scaled volume with 3x3 spatial grids. An example is shown in Fig. 2. To reduce the boundary effect, we also include the 5x5 and 2x2 central overlapped cells in the two scales, respectively. These measurements are concatenated to form an overall appearance descriptor. Also, since the relative location to the scan center is also a useful cue, we include $(dx, dy) = (\frac{|x - scan\_center\_x|}{dim\_x}, \frac{|y - scan\_center\_y|}{dim\_y})$ in our overall descriptor.

Table 1. Statistics of the experimental dataset (ERM: epiretinal membrane, ME: macular edema, AMD: age-related macular degeneration, MH: macular hole). Note that one eye can contain several diseases and may be counted in more than one category.

Num. of Eyes	Normal	ERM	ME	AMD	MH	All diseased	Total Eyes
Training set	30	28	37	16	17	59	89
Testing set	33	19	31	13	15	48	81

Table 2. The localization distance (in pixels) of all methods

Results	Normal (33 cases)		Diseased (48 cases)		Overall (81 cases)	
	mean	median	mean	median	mean	median
Second Expert	1.78±1.37	1.56	1.84±1.42	1.75	1.82±1.39	1.61
S-SVM	2.87±1.45	2.73	3.14±1.96	2.78	3.03±1.77	2.73
B-SVM	3.57±1.94	3.16	3.98±2.15	3.80	3.81±2.06	3.61

Table 3. Percentage of testing scans within various localization distances (in pixels)

Percentage	≤ 2	≤ 4	≤ 6	≤ 8	≤ 10	≤ 12
Second Expert	67.9%	91.4%	98.8%	100%	100%	100%
S-SVM	30.9%	77.8%	95.1%	97.5%	98.8%	100%
B-SVM	18.5%	55.6%	87.7%	97.5%	98.8%	100%

3 Experimental Results

We collected a large sample of 3D SD-OCT macular scans (200x200x1024 or 512x128x1024 protocol, 6x6x2 mm; Cirrus HD-OCT; Carl Zeiss Meditec). Each scan is then normalized to be 200x200x200 in x, y, z. For each scan, two ophthalmologists labeled the (x, y) location of the foveola independently. We then included a total of 170 scans from 170 eyes/126 subjects in which all scans have good expert labeling agreement (distance ≤ 8 pixels). One expert's labeling was adopted as the ground truth while the other was used to assess the inter-expert variability. We split the dataset to a training and a testing set such that they have similar disease distributions, and eyes from the same subject were assigned to the same set. The statistics of our dataset is detailed in Table 1.

We conducted experiments to compare the performance of the proposed S-SVM with binary SVM (B-SVM), both using linear kernel for localization efficiency. We used *SVMStruct* package [13] and *SVMLight* [14] for S-SVM and B-SVM, respectively. The precision ϵ is set to 0.1 and the parameter C is set by performing 2-fold cross validation on the training set. In B-SVM training, for each training scan, we sampled k locations which are at least 8 pixels away from the ground truth as negative examples. We tested for $k = 1, \cdots, 5, 10, 25, 50$. The best result of B-SVM was reported for comparison to S-SVM.

The mean and median localization distance of the S-SVM, B-SVM (best $k =$ 1), and the second human expert are detailed in Table 2. The results for the percentage of scans within various precision are shown in Table 3. From Table 2,

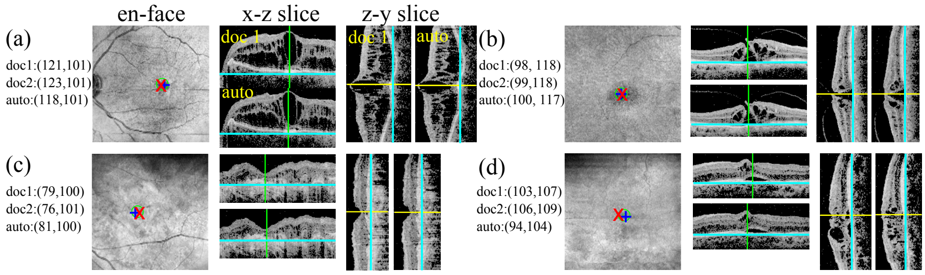


Fig. 3. (a)-(d): example results of the proposed method (auto) compared to the ground truth (doc 1). In the en-face, the labeling of *auto* is marked as "red x", *doc 1* as "green o", and the second expert (doc 2) as "blue +". The slices that cross the foveola defined by *doc 1* and *auto* are shown, where the x, y, z (at RPE layer) position are illustrated in *green*, *yellow*, and *light blue* line. (d): An example of larger error in *auto*.

the performance of the second expert is the best, followed by S-SVM, and then B-SVM. The labeling difference between S-SVM and the second expert is only 1.25 pixels on average, though this is statistically significant (*t-test*, $p \ll 0.001$). From Table 3, our S-SVM can localize 95.1% of scans within 6 pixels, well within the foveola's diameter (12 pixels). Example outputs of S-SVM are in Fig. 3.

In comparison to B-SVM, S-SVM achieved smaller median, mean and standard deviation in all cases as shown in Table 2, and their performance difference is statistically significant (*t-test*, $p = 0.004$). From Table 3, S-SVM also shows larger percentage of scans within anatomical foveola area (95% vs. 87%). S-SVM's better performance is intuitively due to its direct minimization of the localization risk, and its efficient use of all negative locations (the final constraint size $|W| = 22$). In addition, we observed that when using B-SVM, sampling more negative examples (≥ 3 per scan) in training doesn't give us higher performance (some scans have ≥ 20 pixel errors). This is likely due to the higher imbalanced sample number between the two classes that can result in classifier degeneration. Our results demonstrate the value of the proposed S-SVM approach.

The running time of the training of our S-SVM is about 5 hours while for a B-SVM is 1 hour (with 2.67GHz CPU, Matlab+SVM software). Both methods gave the prediction result in 1 minute for each scan. This running time can be improved by parallelizing the score evaluations in sliding window search for both methods, and the loop in finding the most-violated constraint in S-SVM training.

4 Conclusion

In this paper, we propose an effective approach to determine the location of the fovea in retinal 3D-OCT images. Structural SVM is learned to directly predict the foveola location, such that the score at the ground truth position is higher than that of any other position by a margin set to the localization loss. This S-SVM formulation directly minimizes the empirical risk of localization, naturally fitting the localization problem. A coarse-to-fine sliding window search approach is applied to efficiently find the most-violated constraint in the cutting-plane training and in prediction. Our results show that S-SVM outperforms B-SVM, and is within only 1.25 pixel difference on average compared to the second expert.

Our results suggest that the S-SVM paradigm, using the efficient coarse-to-fine sliding window approach during training, could be profitably applied in a broad range of localization problems involving medical image datasets.

Acknowledgments. This research is supported in part by National Institutes of Health contracts R01-EY013178 and P30-EY008098, The Eye and Ear Foundation (Pittsburgh, PA), unrestricted grants from Research to Prevent Blindness, Inc. (New York, NY), and grants from Intel Labs Pittsburgh (Pittsburgh, PA).

References

1. Abramoff, M.D., Garvin, M.K., Sonka, M.: Retinal imaging and image analysis. IEEE Reviews in Biomedical Engineering 3, 169–208 (2010)
2. Yehoshua, Z., Wang, F., Rosenfeld, P.J., Penha, F.M., Feuer, W.J., Gregori, G.: Natural history of drusen morphology in age-related macular degeneration using spectral domain optical coherence tomography. American Academy of Ophthalmology 118(12), 2434–2441 (2011)
3. Liu, Y.Y., Chen, M., Ishikawa, H., Wollstein, G., Schuman, J., Rehg, J.M.: Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding. Medical Image Analysis 15, 748–759 (2011)
4. Niemeijer, M., Abramoff, M.D., van Ginneken, B.: Fast detection of the optic disc and fovea in color fundus photographs. Medical Image Analysis (13), 859–870 (2009)
5. Zhan, Y., Zhou, X.S., Peng, Z., Krishnan, A.: Active Scheduling of Organ Detection and Segmentation in Whole-Body Medical Images. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 313–321. Springer, Heidelberg (2008)
6. Pescia, D., Paragios, N., Chemouny, S.: Automatic detection of liver tumors. In: IEEE Intl. Symposium on Biomedical Imaging (2008)
7. Pedersoli, M., Gonzàlez, J., Bagdanov, A.D., Villanueva, J.J.: Recursive Coarse-to-Fine Localization for Fast Object Detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 280–293. Springer, Heidelberg (2010)
8. Blaschko, M.B., Lampert, C.H.: Learning to Localize Objects with Structured Output Regression. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 2–15. Springer, Heidelberg (2008)
9. Joachims, T., Finley, T., Yu, C.N.J.: Cutting-plane training of strucural SVMs. Journal of Machine Learning (2009)
10. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector learning for interdependent and structured output spaces. In: ICML (2004)
11. Xu, J., Ishikawa, H., Wollstein, G., Schuman, J.S.: 3D OCT eye movement correction based on particle filtering. In: EMBS, pp. 53–56 (2010)
12. Freeman, W.T., Roth, M.: Orientation histogram for hand gesture recognition. In: Intl. Workshop on Automatic Face and Gesture Recognition, pp. 296–301 (1994)
13. Joachims, T.: Support vector machine for complex outputs, software http://svmlight.joachims.org/svm_struct.html
14. Joachims, T.: SVMLight support vector machine, software http://svmlight.joachims.org/