Sliding Window and Regression Based Cup Detection in Digital Fundus Images for Glaucoma Diagnosis*

Yanwu Xu[1], Dong Xu[1], Stephen Lin[2], Jiang Liu[3], Jun Cheng[3], Carol Y. Cheung[4], Tin Aung[4,5], and Tien Yin Wong[4,5]

[1] School of Computer Engineering, Nanyang Technological University, Singapore
[2] Microsoft Research Asia, P.R. China
[3] Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
[4] Singapore Eye Research Institute, Singapore
[5] Department of Ophthalmology, National University of Singapore, Singapore

Abstract. We propose a machine learning framework based on sliding windows for glaucoma diagnosis. In digital fundus photographs, our method automatically localizes the optic cup, which is the primary structural image cue for clinically identifying glaucoma. This localization uses a bundle of sliding windows of different sizes to obtain cup candidates in each disc image, then extracts from each sliding window a new histogram based feature that is learned using a group sparsity constraint. An ϵ-SVR (support vector regression) model based on non-linear radial basis function (RBF) kernels is used to rank each candidate, and final decisions are made with a non-maximal suppression (NMS) method. Tested on the large *ORIGA*$^{-light}$ clinical dataset, the proposed method achieves a 73.2% overlap ratio with manually-labeled ground-truth and a 0.091 absolute cup-to-disc ratio (CDR) error, a simple yet widely used diagnostic measure. The high accuracy of this framework on images from low-cost and widespread digital fundus cameras indicates much promise for developing practical automated/assisted glaucoma diagnosis systems.

1 Introduction

Glaucoma affects about 60 million people [1] and is responsible for approximately 5.2 million cases of blindness (15% of world total) [2]. It unfortunately cannot be cured because the damage to the optic nerve cannot be reversed. Early detection is thus essential for people to seek early treatment and prevent the deterioration of vision [3]. In recent years, much effort has been put into automated/assisted glaucoma diagnosis systems based on computer vision. The design of a glaucoma analysis system depends on the image cues and image modality used.

Among the structural image cues studied for glaucoma diagnosis, those based on the optic disc and cup are of particular importance. The optic disc is located where the ganglion nerve fibers congregate at the retina. The depression inside the optic disc where the fibers leave the retina via the optic nerve head (ONH) is known as the optic cup. The boundaries of the cup and disc structures need to be identified as it facilitates evaluation

* This work is funded by Singapore A*STAR SERC Grant (082 101 0018).

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 1–8, 2011.

of glaucoma cues such as cup and disc asymmetry and large cup-to-disc ratio (CDR), defined as the ratio of the vertical cup diameter to the vertical disc diameter [4]. The CDR value can be determined by planimetry from color fundus images after the optic disc and cup are outlined manually. Since it is very time consuming and labor intensive to manually annotate the cup and disc for each image, computer vision methods have been proposed to automatically segment the disc and cup in fundus images.

In previous work, researchers have mainly focused on automated segmentation of the optic disc [5], using various techniques such as intensity gradient analysis, Hough transforms, template matching, pixel feature classification, vessel geometry analysis, deformable models and level sets [6][7]. In this paper, we focus only on the challenging cup detection problem [8][9], using a large clinical dataset called *ORIGA*$^{-light}$ [10] in which the ground-truth of discs and cups is marked by a team of graders from a hospital. Unlike previous segmentation based algorithms, which classify each pixel as cup or non-cup, our technique identifies a cup as a whole, based on sliding windows and machine learning.

2 Sliding Window Based Cup Detection

In this work, we start with a disc image for cup detection, which may be obtained using methods such as [6]. Different from previous image processing based techniques, a general sliding window based learning framework is proposed for cup localization.

2.1 Sliding Windows

From the suggestion of doctors and graders, in this paper we represent the localized disc by a non-rotated, arbitrary-sized ellipse denoted by its central point (U, V), corresponding description function $\frac{(x-U)^2}{U^2} + \frac{(y-V)^2}{V^2} = 1$, and rectangular bounding box delimited by $(1, 1)$ and $(2U - 1, 2V - 1)$. With the disc image, we search for the candidate cup by sampling non-rotated ellipses at various aspect ratios represented as $(\mathbf{u}, \mathbf{v}, \mathbf{r}, \mathbf{s})_{N_w \times 4}$, where $(\mathbf{u}, \mathbf{v}, \mathbf{r}, \mathbf{s})$ is the description matrix of all the cup candidates and N_w is the number of cups. For the i^{th} cup candidate denoted as (u_i, v_i, r_i, s_i), its description function is $\frac{(x-u_i)^2}{r_i^2} + \frac{(y-v_i)^2}{s_i^2} = 1$ and $|r_i| + |u_i| \leq |U|$, $|s_i| + |v_i| \leq |V|$. Cup candidates are generated by sampling values of $(p_i^u, p_i^v, p_i^r, p_i^s)$. In this work, we empirically set $(u_i, v_i, r_i, s_i) = (U \cdot p_i^u, V \cdot p_i^v, U \cdot p_i^r, V \cdot p_i^s)$, where $p_i^u \in [0.75, 1.25]$, $p_i^v \in [0.75, 1.25]$, $p_i^r \in [0.2, 1]$ and $p_i^s \in [0.2, 1]$ with a sampling interval of 0.06. In the detection phase, with this setting, for the input discs with different sizes, $N_w = 6691$ cup candidates from each disc image can be obtained with the same sampling values of $\{(p_i^u, p_i^v, p_i^r, p_i^s)|_{i=1}^{N_w}\}$.

2.2 Feature Representation

Features play an important role in computer vision applications. In this paper, we introduce a new region based color feature for cup detection. Similar to segmentation based approaches, it takes advantage of color differences between cup and disc regions in fundus images. However, it additionally accounts for the elliptical shape of a cup and the blood vessels that run through it, which often mislead segmentation algorithms. We extract features using the following steps:

1. For a given disc image, the green, blue, hue and saturation channels are computed from the color image. Since the red (RGB model) and value (HSV model) channels differ little between the disc and cup, they are not used in this work. We linearly scale hue and saturation values into [0,255] for consistency with the green and blue color channels. For each color channel, its values are histogrammed by quantization with different bin numbers $\mathscr{B} = \{B_n|_{n=1}^{N}\}$ such that each bin has an equal (or as equal as possible due to quantization) number of pixels, giving equalized channels. In the experiments, we use $\mathscr{B} = \{3, 4, \cdots, 9, 10, 12, 16, \cdots, 28, 32\}$.
2. For each color channel and each number of bins $B_n \in \mathscr{B}$, we form three types of features: 1) L1 normalized histogram of the candidate cup region; 2) L1 normalized histogram of the candidate non-cup region within the disc; 3) for each of the B_n bins, the proportion of cup pixels with respect to all the pixels within the disc. Determining the optimal bin numbers in each color channel is non-trivial, so we used multiple bin numbers to generate redundant features and then employ a group sparsity based approach to select the most effective and discriminant features. Finally, each feature is represented as a B_i dimensional vector, and we refer to each type of feature for a given color channel and bin number as a *group*.
3. For a candidate cup in a specific disc, referred to as a “cup-disc” candidate, its original feature $\mathbf{f}_i$ is obtained by concatenating 3 types of features over 4 color channels and multiple bin numbers. In our experimental setting, this leads to a feature dimension of $|\mathbf{f}_i| = \sum_{n=1}^{N} 3 \times 4 \times B_n = 12 \sum_{n=1}^{N} B_n = 2208$.

As illustrated in Fig. 1, after the green channel image is histogrammed into three bins, the first bin (illustrated as black pixels) occupies most of the vessel region, the second bin (grey color) mainly occupies the non-cup region, while the third bin (white color) occupies most of the cup region. Also, it can be observed that the equalized channels are more clear and they facilitate distinguishing different components, since they are relatively insensitive to illumination condition (*e.g.*, see the hue channel). For the cup detection task, it is unclear which color channels to use and how many bins is optimal for a given channel, so we apply statistical learning methods to select features from this large redundant feature representation and use only the selected features for cup localization.

2.3 Feature Selection Based on Group Sparsity Constraint

Identifying and using only the effective elements of the original feature can bring higher precision and speed. For a cup candidate in the training set with an original feature $\mathbf{f}_i$ consisting of g feature groups, we denote its regression value (*i.e.*, the score obtained from its overlap ratio with the clinical ground-truth) as $z_i \in [0, 1]$. We adopt the linear regression model $\omega^T \mathbf{f}_i + \mu$ to obtain the estimated value, where ω is the weighting vector and μ is the bias. We minimize the following objective function:

$$\min_{\omega,\mu} \sum_{i=1}^{l} \|z_i - \omega^T \mathbf{f}_i - \mu\|^2 + \lambda \sum_{j=1}^{g} \|\omega_j\|_2 \qquad (1)$$

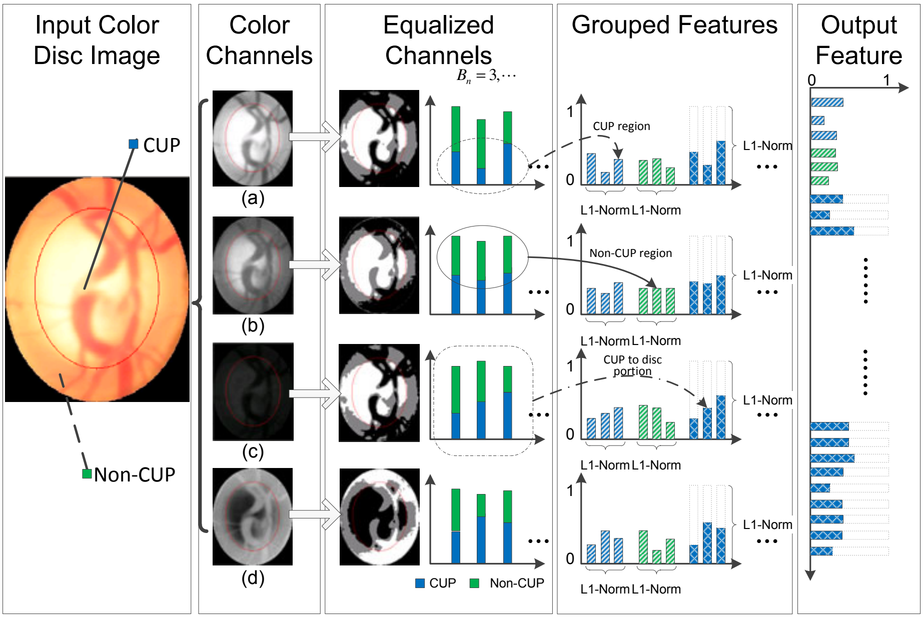

Fig. 1. Grouped feature extraction for cup localization. (a) (b) (c) and (d) represent green, blue, hue and saturation color channels, respectively.

where ω_j is the corresponding weight of the j^{th} feature group, l is the number of training samples and λ is used to control the sparsity of ω. In Eq. (1), the first term represents the regression error and the second term is a $L_{1,2}$-norm based regularizer to enforce group sparsity. Considering the features are intrinsically organized in groups, we use an $L_{1,2}$-norm based regularizer to select features from only a sparse set of groups. In the experiments, we use the group-lasso method in [11] to solve Eq. (1).

After ω is obtained, it can be used as a feature selection mask to generate the final features, *i.e.*, the j^{th} group of features is selected when $\|\omega_j\|_2 > 0$. We represent the feature extracted from the i^{th} cup-disc training sample after feature selection as $\mathbf{x}_i$. The lower dimension of the final feature $\mathbf{x}_i$ leads to faster feature extraction and cup detection in the testing phase when compared with using the original 2208-D feature.

2.4 Non-linear Regression Model

After feature selection, we introduce a kernelized ϵ-SVR to further improve accuracy:

$$\min_{\mathbf{w},\gamma,\xi,\xi^*} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \quad s.t. \quad \begin{cases} \mathbf{w}^T\phi(\mathbf{x}_i) + b - z_i \leq \epsilon + \xi_i, \\ z_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, i = 1, \cdots, l \end{cases} \tag{2}$$

where $\mathbf{x}_i$ is a training sample after feature selection, ξ_i and ξ_i^* are slack variables for ϵ-insensitive loss, C is a regularization parameter, $\mathbf{w}^T\phi(\mathbf{x}_i) + b$ is the non-linear regression function with $\mathbf{w}$ as the weight vector in the feature space, b as the bias term, and $\phi(\cdot)$ is the non-linear function mapping $\mathbf{x}_i$ from the original space to a higher dimensional space. LibSVM toolbox [12] is used to solve this problem in our implementation.

In the testing phase, the feature $\mathbf{x}_i$ is extracted directly from the i^{th} cup candidate ($i = 1, 2, \cdots, N_w$) in the test disc image based on the feature selection mask ω. Then the regression values of all the cup candidates are calculated, denoted as $\gamma = (\gamma_1, \cdots, \gamma_i, \cdots, \gamma_{N_w})^T$. We sort γ in descending order and obtain the final detection result using the non-maximal suppression (NMS) of the next section.

2.5 Detection Result Fusion with NMS

Various NMS methods have been proposed to reduce redundancy in sliding window based detection. Let us denote the cup candidates as $\mathscr{D} = \{D_1, D_2, \cdots, D_{N_w}\}$, where D_i is represented as (u_i, v_i, r_i, s_i). Note that the cup candidates are sorted according to the regression value γ_i. A detection result can simply be computed as the mean of the top T candidates with the highest regression values, $\overline{D}_T : (\overline{u}_T, \overline{v}_T, \overline{r}_T, \overline{s}_T)$.

Since the top T candidates $D_i|_{i=1}^T$ may not all be of high accuracy, we perform the following steps to handle outliers, similar to majority voting:

1. Initialize a zero matrix $O_{(2U-1)\times(2V-1)}$ of the same size as the disc image.
2. For each cup candidate $D_i|_{i=1}^T$, add a vote for each pixel that lies within D_i.
3. Locate the minimal rectangular bounding box $B_{NMS} : (E_l, E_r, E_t, E_b)$ containing the pixels with no fewer than $\rho \cdot T$ votes, where E_l, E_r, E_t and E_b represent the left, right, top and bottom bounds, respectively, and ρ is a threshold empirically fixed to 0.75 in this work.
4. The final detected cup is represented by the ellipse: $(\frac{E_r+E_l}{2}, \frac{E_t+E_b}{2}, \frac{E_r-E_l+1}{2}, \frac{E_t-E_b+1}{2})$.

3 Experiments

In this section, we describe the evaluation criteria and experimental setting, then analyze the two main steps in our framework, *i.e.*, the group sparsity based feature selection and candidate cup ranking by using RBF based ϵ-SVR, through comparisons of three cup detection methods. The first method (referred to as feature selection+ϵ-SVR) uses the group sparsity based feature selection method to obtain a low-dimensional feature and then performs RBF based ϵ-SVR to rank the cup candidates. The second method (referred to as feature selection+simple ranking) uses $\omega^T \mathbf{f}_i$ to directly rank the cup candidates after obtaining ω from feature selection. In the third method (referred to as ϵ-SVR), we directly perform RBF based ϵ-SVR ranking using the original feature $\mathbf{f}_i$ without conducting the feature selection process. We also compare our feature selection+ϵ-SVR approach with level-set based segmentation methods [6][9].

3.1 Cup Detection Evaluation Criteria

Three evaluation criteria are commonly used for cup detection/segmentation, namely non-overlap ratio (m_1), relative absolute area difference (m_2) [13] and absolute cup-to-disc ratio (CDR) error (δ), defined as:

$$m_1 = 1 - \frac{area(E_{dt} \bigcap E_{gt})}{area(E_{dt} \bigcup E_{gt})},\ m_2 = \frac{|area(E_{dt}) - area(E_{gt})|}{area(E_{gt})},\ \delta = \frac{|d_{dt} - d_{gt}|}{R} \quad (3)$$

where E_{dt} denotes a detected cup region, E_{gt} denotes the ground-truth ellipse, d_{dt} is the vertical diameter of the detected cup, d_{gt} is the vertical diameter of the ground-truth cup, $R = 2V - 1$ is the vertical diameter of the disc, and $0 < d_{dt}, d_{gt} \leq R$.

3.2 Experimental Setup

Training samples. The *ORIGA*$^{-light}$ dataset is divided into two sets S_A and S_B, which consist of 150 images and 175 images, respectively. In the training phase, 500 samples including one ground-truth cup and 499 randomly generated cup candidates are obtained for each of the 150 disc images from set S_A. In total, we have 75,000 cup candidates in the training set. The method for generating training cup candidates in the training phase is designed so that the windows of the training cup candidates and those examined in the testing phase have different description parameters (u, v, r, s). We then use both image sets for testing our algorithm.

Parameter setting for feature selection. For each cup-disc candidate, its original feature $\mathbf{f}_i$ is extracted, and the regression value corresponding to the overlap ratio $(1 - m_1)$ of the cup candidate region and the ground-truth ellipse is also calculated using Eq. (3). Only the ground-truth cup region will have a full score of 1. We solve the problem in Eq. (1) using the group-lasso tool [11] to obtain ω by empirically setting the parameter $\lambda = 0.01$. According to the obtained values of $||\omega_j||_2$, 993 of 2208 feature dimensions are selected. Using only 44.97% of the original features leads to significant acceleration in detection speed.

Parameter setting for RBF based ϵ-SVR. The well-known Lib-SVM toolbox [12] is used to train the ϵ-SVR model. We perform cross-validation to determine the optimal parameters by setting the parameters as $C \in \{10^{-3}, 10^{-2}, \cdots, 10^2, 10^3\}$, $\epsilon \in \{10^{-3}, 10^{-2}, 10^{-1}\}, p \in \{10^{-3}, 10^{-2}\}$, and $g = 2^k \cdot \frac{1}{2\sigma^2}$ with $k \in \{-7, -5, \cdots, 5, 7\}$, where p is the convergence threshold in the training phase and σ^2 is the mean of all the Euclidean distances between any two training samples. The samples $\mathbf{x}_i|_{i=1}^{l}$ are obtained by applying the feature selection mask ω onto the original features $\mathbf{f}_i|_{i=1}^{l}$. To avoid overlap between the training and testing samples in the cross-validation process, 8000 randomly selected samples and the ground-truth cups from the first 100 images are used for training, while another 6000 randomly selected samples and the ground-truth cups from the remaining 50 images are used for testing. After conducting cross-validation, the optimal parameters were determined to be $C = 10$, $p = 10^{-3}$, $k = -3$ and $\epsilon = 0.001$. With these parameters, all of the 75,000 samples are used to train an ϵ-SVR model for the testing phase.

3.3 Comparison of Three Methods in Our Framework

We compared the three methods to show the effectiveness of each step of our framework. The same cross-validation method is used to determine the optimal parameters of the ϵ-SVR method for a fair comparison. The results are listed in Table 1. From it, we have the following observations:

Table 1. Comparison of three methods in our framework and level-set based methods

Method	*Set* S_A			*Set* S_B			$S_A\&S_B$		
Evaluation criteria	m_1	m_2	δ	m_1	m_2	δ	m_1	m_2	δ
Feature Selection+ϵ-SVR	**0.254**	**0.252**	**0.081**	**0.289**	**0.409**	**0.106**	**0.268**	**0.315**	**0.091**
Feature Sel.+Simple ranking	0.301	0.398	0.115	0.344	0.643	0.143	0.324	0.530	0.130
ϵ-SVR	0.269	0.314	0.101	0.320	0.484	0.128	0.290	0.382	0.112
Level-set [6]	0.458	0.625	0.137	0.552	1.189	0.214	0.495	0.847	0.162
Level-set+Hist-analysis [9]	0.458	0.519	0.119	0.491	0.859	0.159	0.476	0.702	0.140
Relative error reduction to [9]	44.5%	51.5%	31.9%	41.1%	52.4%	33.3%	**43.7%**	**55.1%**	**35.0%**

1. Comparing feature selection+ϵ-SVR to feature selection+simple ranking shows that the RBF kernel based ϵ-SVR is better than simple ranking using the selected features from our feature selection method. This demonstrates better generalization ability of ϵ-SVR, which is consistent with previous work on image classification.
2. Comparing feature selection+ϵ-SVR to ϵ-SVR shows that group sparsity based feature selection also improves performance by selecting and using the most effective and discriminant features. Moreover, it accelerates the detection procedure by about 60%. We also observe that the performance improvement from feature selection+ϵ-SVR over ϵ-SVR is not as large as that from feature selection+ϵ-SVR over feature selection+simple ranking, possibly because ϵ-SVR also tunes the weight of each feature dimension and thus acts as a kind of feature selection.

3.4 Comparison with Level-Set Based Segmentation [6],[9]

One of the few methods for both cup and disc segmentation is the level-set method of [6], which first identifies the pixels that belong to the cup region, then uses a convex hull method to generate an ellipse. In [9], histogram based analysis of the color pixel intensity together with multiple method fusion are also employed to further improve cup detection accuracy. Table 1 compares our method to these two level-set approaches[1]. Compared with the more advanced approach [9], our method is shown to significantly improve cup localization accuracy in both sets S_A and S_B, and m_1 and CDR error (i.e., δ) are reduced by 43.7% and 35.0%, respectively. We note that all methods obtain better performance on set S_A, possibly because of the data distribution itself. Moreover, it is worth mentioning that the relative CDR error reduction in set S_B is more significant when compared with that in set S_A.

3.5 Detection Speed and Limitations

The experiments were carried on an eight-core $2.67GHz$ PC with $16GB$ RAM using the Matlab Parallel Computing Toolbox. In our approach, the extraction of the original feature takes about 6 minutes per image, while feature selection reduces the time

[1] We did not compare with the Haar+Adaboost method for general object detection, because the cup detection task was formulated as a regression problem, not a binary classification problem, and the Haar feature is not suitable for objects with varying aspect ratios.

cost by about 60%. The RBF based ϵ-SVR takes about 1 minute per image. The NMS takes about 0.2 minutes per image on average. The main time cost is for feature extraction, and the proposed sparsity based feature selection greatly accelerates the detection speed. In addition, from our observations the proposed method does not handle large cups as effectively, because NMS suppresses the rim of the cup.

4 Conclusion

We proposed a sliding window based learning framework with a newly developed feature for cup detection in glaucoma diagnosis. Tested on a large clinical dataset with three evaluation criteria, it achieves a 26.8% non-overlap ratio (m_1) with manually-labeled ground-truth, a 31.5% relative absolute area difference (m_2) and a 0.091 absolute CDR error (δ). In future work, we plan to elevate performance using new features or by introducing domain-specific knowledge on this problem.

References

1. Wong, T., Loon, S., Saw, S.: The epidemiology of age related eye diseases in asia. British Journal of Ophthalmology 90(4), 506–511 (2006)
2. Thylefors, B., Negrel, A.: The global impact of glaucoma. Bull. World Health Organ. 72(3), 323–326 (2006)
3. Michelson, G., Wärntges, S., Hornegger, J., Lausen, B.: The papilla as screening parameter for early diagnosis of glaucoma. Dtsch. Arztebl. Int. 105(34-35), 583–589 (2008)
4. Jonas, J., Budde, W., Panda-Jonas, S.: Ophthalmoscopic evaluation of the optic nerve head. Survey of Ophthalmology 43, 293–320 (1999)
5. Abramoff, M., Alward, W., Greenlee, E., Shuba, L., Kim, C., Fingert, J., Kwon, Y.: Automated segmentation of the optic disc from stereo color photographs using physiologically plausible features. Investigative Ophthalmology and Vis. Sci. 48(4), 1665–1673 (2007)
6. Liu, J., Wong, D., Lim, J., Li, H., Tan, N., Zhang, Z., Wong, T., Lavanya, R.: ARGALI: an automatic cup-to-disc ratio measurement system for glaucoma analysis using level-set image processing. In: Int. Conf. Biomed. Eng. (2008)
7. Li, C., Xu, C., Gui, C., Fox, M.: Level set evolution without re-initialization: A new variational formulation. In: IEEE Conf. Compt. Vis. Patt. Rec., pp. 430–436 (2005)
8. Merickel, M., Wu, X., Sonka, M., Abramoff, M.: Optimal segmentation of the optic nerve head from stereo retinal images. In: Medical Imaging: Physiology, Function, and Structure from Medical Images, vol. 6143 (2006)
9. Wong, D., Liu, J., Lim, J., Tan, N., Zhang, Z., Lu, S., Li, H., Teo, M., Chan, K., Wong, T.: Intelligent fusion of cup-to-disc ratio determination methods for glaucoma detection in ARGALI. In: Int. Conf. Engin. in Med. and Biol. Soc., pp. 5777–5780 (2009)
10. Zhang, Z., Yin, F., Liu, J., Wong, D., Tan, N., Lee, B., Cheng, J., Wong, T.: Origa-light: An online retinal fundus image database for glaucoma analysis and research. In: Int. Conf. IEEE Engin. in Med. and Biol. Soc., vol. 2010, pp. 3065–3068 (2010)
11. Jacob, L., Obozinski, G., Vert, J.P.: Group lasso with overlap and graph lasso. In: Int. Conf. Machine Learning. ACM, New York (2009)
12. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. ACM Trans. on Intel. Sys. and Tech. 2(3), 27:1–27:27 (2011)
13. Deng, X., Du, G.: Liver tumor segmentation. In: 3D Segmentation in the Clinic: A Grand Challenge II Workshop at 10th MICCAI (2007)