A HYBRID FEATURE SELECTION STRATEGY FOR IMAGE DEFINING FEATURES: TOWARDS INTERPRETATION OF OPTIC NERVE IMAGES

[1]JIN YU, [1]SYED SIBTE RAZA ABIDI, [2]PAUL HABIB ARTES

[1]Faculty of Computer Science, Dalhousie University, Halifax B3H 1W5, Canada
[2]Deptartment of Ophthalmology and Visual Sciences, Dalhousie University, Halifax B3H 1W5, Canada
E-MAIL: sraza@cs.dal.ca

Abstract:

Modern imaging techniques such as Confocal Scanning Laser Tomography (CSLT) capture high-quality optic nerve images. The automated analysis of CSLT images, by combining image processing and data mining methods, offers the potential for developing objective methods for supporting clinical decision-making in glaucoma. We present our approach that involves the analysis of CSLT images using moment methods to derive abstract image defining features, and then the use of these features to train classifiers for automatically distinguishing CSLT images of healthy and diseased optic nerves. As a first step, in this paper, we present investigations in feature subset selection methods for reducing the relatively large input space produced by the moment methods. Our results demonstrate that our methods can discriminate between healthy and glaucomatous optic nerves based on shape information automatically derived from CSLT tomography images.

Keywords:

Feature Selection; Optic Nerve Images; Zernike Moments; Markov Blanket; Confocal Scanning Laser Tomography.

1. Introduction

Confocal Scanning Laser Tomography (CSLT), a modern eye imaging technique, captures 3-dimensional optic nerve images that can be analyzed, in an automatic manner, to provide support in the clinical care of glaucoma patients [1]. Yet, to date, most diagnostic tools require human intervention—a trained professional has to manually define the margins of the optic nerve (a process that is somewhat subjective in nature and highly dependent on training and expertise). Whilst CSLT image analysis has tremendous potential to improve the clinical care of glaucoma patients, current methods for image analysis fail to detect optic nerve damage sufficient accuracy [2].

In this on-going project, we are working towards the development of a data-driven glaucoma diagnostic support system (shown in figure 1) that features the automatic interpretation of CSLT images by (a) applying image processing techniques to derive image-defining data that can be applied to a suite of data mining algorithms; (b) selecting a subset of image features that exhibit optimal classification capabilities for distinguishing between healthy and diseased optic nerves, and between different subtypes of optic nerve damage; (c) inducing classification rules in order to provide domain experts a symbolic explication of the data and the inherent class structures.

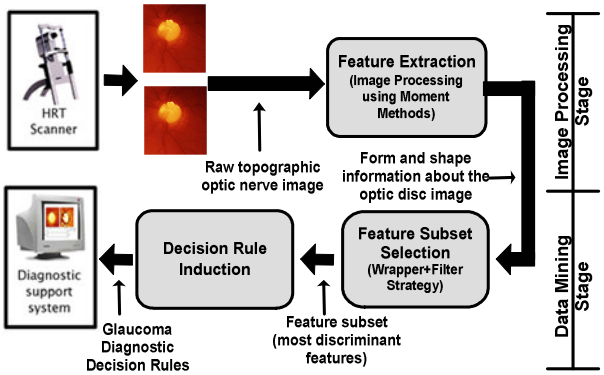

Figure 1. Functional design of a glaucoma diagnostic support system

In this paper we present an automated approach to CSLT image analysis by using image processing methods to derive image-defining features, and then applying data mining methods to the derived features to derive diagnostic knowledge from the images. The data mining task is the classification of CSLT images—to discriminate between healthy and glaucomatous optic nerves—based on the optic nerve's shape. However, given the large number of features produced by image processing methods, it is important to reduce the feature subset size, without losing information, to

optimize the image classification task.

First we present, the derivation of image-defining features from CSLT images using Moment Methods [3, 4]. Classification of CSLT images based on image features (or moments) is constrained by the relatively large input space—i.e. image features—produced by moment methods, thus prompting the need to applying feature selection methods [5, 6, 7, 8] to select a feature subset that offers optimal classification accuracy for classifying CSLT images of normal and glaucoma patients. We have developed a two-pass feature subset selection method that is a hybrid of wrapper and filter methods. In the first pass, wrapper models of Multilayer Perceptron (MLP) [9] and Support Vector Machines (SVM) are used in a forward feature selection manner to identify an optimal subset of lower order image-defining moments that offer optimal classifications. In the second pass, the Markov blanket filter method [10] is used to select the highly relevant moments/features from the feature subset selected in pass 1. At the completion of the two feature selection passes we identify the smallest possible set of moments/features that provide the highest classification accuracy. Our results will demonstrate the efficacy of our automated approach to discriminate between healthy and glaucomatous optic nerves, based on shape information derived from CSLT topography images.

Analysis of optic nerve data and CSLT based images, particularly using an assortment of feature subset selection and data classification methods has been actively pursued by researchers, with varying results [11-15]. Bowd et al [11], working with retinal tomograph images applied forward and backward feature selection methods for training MLP, SVM and linear discriminant functions; Park et al [12] have used correlation analysis and forward wrapper model to select features from optic nerve data for training SVM classifiers; Swindale et al [13] used a hill climbing wrapper model for feature selection to train SVM classifiers; whereas Cheng et al [14] and Peters et al [15] did not apply feature selection prior to their respective image analysis methods.

2. Optic Nerve Image Processing

The Heidelberg Retina Tomograph (HRT) is a CSLT system that uses a low-intensity monochromatic laser beam to scan the back of the eye sequentially in two dimensions to acquire a series of images from consecutive focal planes. Within each image series, the relative height of the retinal surface structure can be inferred by finding the focal plane in which maximum reflectance of each pixel occurs (topography image). After several image series for an eye are acquired (as shown in figure 2), the final mean topography images are used for diagnosis [1].

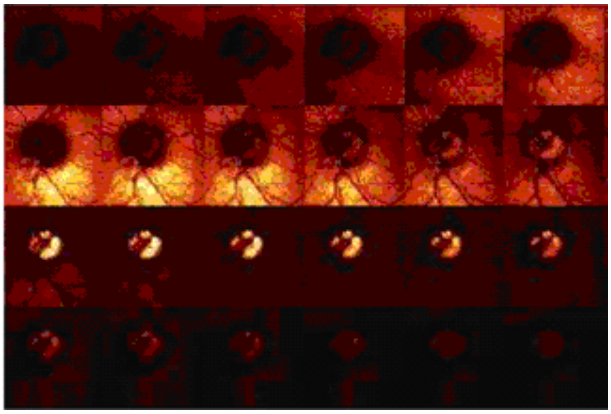

Figure 2. CSLT image series of the optic disc

We use an image processing technique, referred to as Moment Methods [3], to extract features from CSLT images. By describing the properties of connected regions in binary images, Moment features are invariant to translation, rotation and scale. Thus, moment features both describe the image content with respect to its axes and capture detailed geometric information about the image.

In our work, we analyze CSLT images using Zernike moments [4] which use a set of complex polynomials to form a complete orthogonal basis set on the unit disc $(x^2 + y^2) <= 1$ (where x and y define the origin of the pixel). Put simply, Zernike moments describe the image's properties by their order (n) and repetition (m) with respect to a digital image—the low order moments capture gross shape information and high order moments incrementally resolve high frequency information (representing detail) of the digital image. Two attractive features of this analysis is that (a) moments can be made invariant to shifts, rotations and magnification changes; and (b) the optic nerve is centered in the image, thus avoiding the requirement for an independent segmentation stage in which the object is explicitly identified.

It should be noted that typically the low order moments capture fundamental geometric properties and high order moments represent detailed information of the image [4]. However, for image classification based on gross shape it can be argued that the high order moments do not contribute much information; in fact they can be regarded as noise. Given the above assumption, to better classify CSLT images between normal and glaucoma, the primary task is select an optimal number of lower order moments. However, the problem faced is two-fold: (a) there is no available objective measure to determine the exact number of (low order) moments necessary for achieving high classification accuracy; and (b) there is no discernable relationship

between the moments that can be utilized. Hence, there is a need for a feature selection strategy to objectively select an optimal set of moments, starting from the lowest order moments and moving towards higher order moments.

3. CSLT Optic Nerve Data

For our experiments we worked with 1257 tomography images taken at different time intervals from 136 subjects (51 healthy subjects and 85 glaucoma patients). For each CSLT image we generated 254 Zernike moments with order 1 to 29. For the Zernike moments generated, the order n and repetition m meet the conditions n-$|m|$ = *even* and $|m| <= |n|$.

Given the set of 254 moments for each CSLT image, the objective is to determine a set of optimal moments that can provide high classification accuracy. The rationale for feature subset selection is based on the observation that a large number of abstract moments tend to compromise the accuracy of supervised learning classifiers, the classification rules are difficult to understand and the computational cost is high.

4. Hybrid Feature Subset Selection Strategy

Feature selection, namely feature subset selection, is to find a subset of the original features of a data set such that an induction algorithm applied to the selected feature subset generates a classifier with the highest possible accuracy [5].

We have developed a hybrid feature subset selection strategy that combines both wrapper and filter models of feature subset selection, and operates in two phases (illustrated in figure 3). In the first phase, MLP and SVM based wrapper models are used to find an *Optimal Moment Feature Subset* (OMFS) which is the set consisting of low order moment feature groups that provide optimal data classification accuracy. In the second phase, a filter model based on a Markov Blanket (of the class label) [10] is applied to an inferred Bayesian network based on the OMFS. The moments that have no causal relationship with the class are removed from OMFS to realize an even smaller feature sub-set of moments.

4.1. Phase I: Using MLP and SVM

In the absence of any guiding principle to determine the size of the OMFS, we devised an accumulative feature selection strategy, whereby we incrementally add moments to an existing feature set and train a classifier (MLP and SVM) to determine the classification accuracy for the new feature subset. We had two options to generate the feature subset for training: (i) to add the next N higher features to the existing data-set, where N was deemed to range between 1-10 moments. Say, if N = 5, then feature subset1 would include moments 1-5, feature subset2 would add the next 5 moments to contain moments 1-10 and so on; or (ii) to use the intrinsic partitioning of the moments based on their order; the 254 moments were divided into 29 groups based on their order ranging from 1 to 29. This implies that feature subset1 includes moments with *order2*, feature subset2 includes moments with order 2+3, and so on. Both from a theoretical and experimental point of view, we determined that generating the OMFS based on adding moments of increasing order is a sound option.

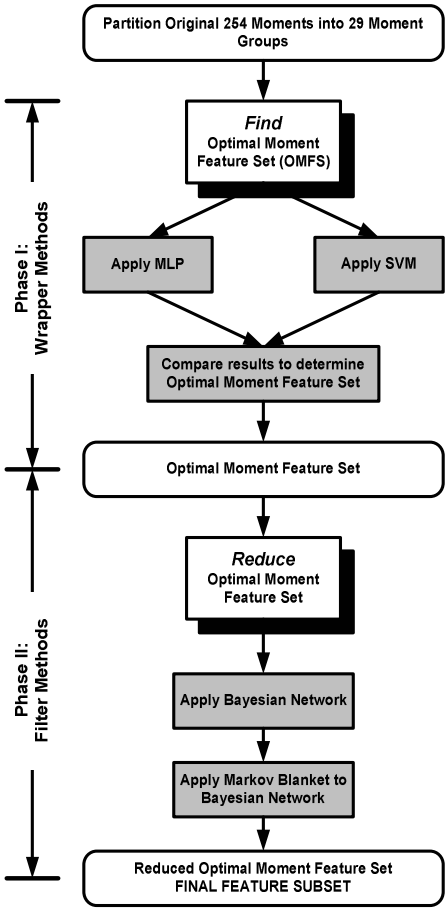

Figure 3. Our Feature Subset Selection Strategy

Finally, to determine the size of the OMFS we generate 29 different feature subsets (in an accumulative manner), and for each feature subset we train two classifiers—MLP and SVM—and determine their classification accuracy. The classification accuracy trend for each of the 29 classifiers is

plotted; the point on the plot (i.e. the moment group) from which the classification accuracy takes a downward trend (with the inclusion of the next higher moment group) is determined as the OMFS.

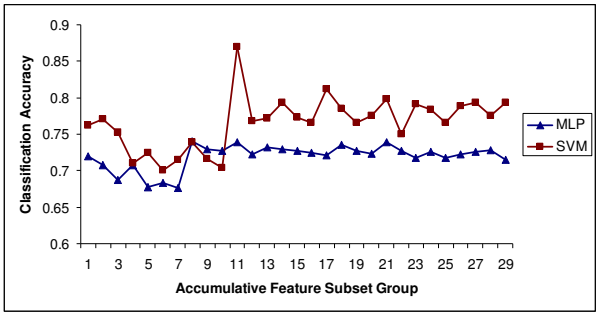

Figure 4. Classification accuracy for both MLP and SVM

Table 1. Classification accuracy and standard deviation for MLP and SVM.

	MLP		SVM	
	Accuracy	**SD**	**Accuracy**	**SD**
1	0.7197	0.0737	0.7618	0.0572
2	0.7071	0.0735	0.7706	0.0724
3	0.6868	0.0760	0.7529	0.0740
4	0.7072	0.0661	0.7103	0.0823
5	0.6769	0.0702	0.7250	0.0818
6	0.6832	0.0730	0.7000	0.0736
7	0.6762	0.0852	0.7147	0.0733
8	**0.7400**	0.0685	0.7397	0.0590
9	0.7297	0.0680	0.7162	0.0679
10	0.7271	0.0609	0.7044	0.0914
11	0.7393	0.0752	**0.8696**	0.0305
12	0.7224	0.0668	0.7676	0.0464
13	0.7324	0.0690	0.7721	0.0482
14	0.7294	0.0647	0.7941	0.0446
15	0.7268	0.0700	0.7735	0.0427
16	0.7241	0.0829	0.7662	0.0661
17	0.7210	0.0708	0.8117	0.0746
18	0.7359	0.0713	0.7853	0.0652
19	0.7272	0.0723	0.7662	0.0735
20	0.7235	0.0925	0.7750	0.0751
21	0.7390	0.0756	07985	0.0632
22	0.7271	0.0961	0.7500	0.0667
23	0.7170	0.0789	0.7912	0.0541
24	0.7257	0.0726	0.7838	0.0632
25	0.7176	0.0713	0.7662	0.0792
26	0.7224	0.0874	0.7882	0.0562
27	0.7257	0.0718	0.7941	0.0581
28	0.7288	0.0723	0.7750	0.0659
29	0.7144	0.0740	0.7941	0.0631

For training the MLP we partitioned the feature subset—i.e. the data—into a training and test set. Different data partitions, ranging from 80%-20%, 75%-25% and 70%-30% (training%-testing%), were used. The classification accuracy results for MLP for all the moment groups are given in table 1. Figures 4 plots the classification trend for MLP.

For training the SVM, each candidate feature subset—i.e. the data—was divided into 75% training and 25% testing set. Based on the training data, a 5-fold cross validation was performed to find the optimal hyper parameters: C and λ. Finally, the testing data was used to calculate the SVM's classification accuracy. In order to minimize the stochastic nature of the method, each candidate feature subset was trained 20 times and the mean classification accuracy is regarded as the final accuracy (as shown in table 1 and figure 4).

A comparison of the classification accuracy trends for both the MLP and the SVM classifiers (see table 1 and figure 4) shows that both these classifiers have a similar classification accuracy trend—i.e. they both start with a relatively high accuracy with the first moment group and then the accuracy drops with the accumulation of the next few moment groups. But later the accuracy starts to pick up again such that for the MLP it peaks when the feature subset constitutes the first 8 moment groups, whereas for the SVM the accuracy peaks for the first 11 moment groups. Furthermore, the classification accuracy with higher order moment groups is relatively low as compared to the peak achieved with the lower order moments.

Based on the above interpretation of the classification accuracy trend for both classifiers, we determined the OMFS to constitute the first 11 moment groups—i.e. the first 47 moment features—because the SVM exhibited the highest accuracy and the MLP exhibited the second highest accuracy with the first 11 moment groups

4.2. Stage II: Using Markov Blanket

Stage I generates the OMFS, which in this case comprises the first 11 groups of moments totaling 47 features. In Stage II, the OMFS is further reduced by selecting highly salient features using a filter model based on a Bayesian network and the Markov blanket of the class label [10]. The choice of Markov blanket is guided by the observation that the correlation between most of moment features and class label is found to be weak, and the same is true for correlation between different features. Hence, correlation based feature selection methods are not suitable. The use of common forward or backward feature selection method is not suitable for our data because when all moments are ordered by Pearson coefficients with class label beginning with the highest one, the classification vary

greatly with the inclusion of the next moment feature. Hence, we decided to use Markov blanket approach as it considers every feature's probability dependence relationship during the learning procedure of the Bayesian network's structure.

A Bayesian network is a directed acyclic graph, where each node represents a random variable and each arc represents a probabilistic dependence. In a Bayesian network where C_A is the set of children of node A, and Q_A is the set of parents of node A, the subset of nodes containing Q_A, C_A and the parents of C_A is called Markov blanket of A. From the above, the Markov blanket of a specific feature is a subset of nodes in the Bayesian network; it comprises the feature's parent nodes, child nodes and all parent nodes of the child nodes. If we consider the class label node as the root node to learn a Bayesian network from data, then all nodes within the Markov blanket of the class node have probabilistic dependence relationship with it. So the Markov blanket method can be used as the criterion for feature selection.

The following steps generated the Markov blanket.

Step 1: We use the *K*2 algorithm to learn the Bayesian network. Initially, the 47 features in the OMFS are discretized using an entropy-based method. As a result, 29 features are discretized into a single value. According to the principle of *K*2 algorithm, all these 29 features are removed from the data-set and only the remaining 18 features are kept for learning the Bayesian network. They following moments were retained: moments {1, 2, 5, 6, 7, 12, 16, 21, 23, 25, 27, 33, 36, 37, 43, 44, 45, 46}.

Step 2: A Bayesian network is learnt based on 18 features, in their original order, using 5-folds stratified cross validation to evaluate the classification accuracy. The resulting classification accuracy was found to be 77.21%.

Step 3: All features were ordered according to the chi squared statistical test score x^2 between the features and the class labels beginning with the highest x^2 score. The moments were ordered as follows: moments {1, 43, 16, 25, 21, 23, 6, 5, 36, 2, 27, 33, 37, 7, 46, 45, 44, 12}.

Step 4: To test the correctness of ordered moments, a Bayesian classier was learnt that gave a classification accuracy of 80.88%, which is 3% higher than that based on the original order of the features. So ordering features using chi square score improved the Bayesian classifier's accuracy. Figure 4, illustrates the learnt Bayesian network.

Step 5: We know that the Markov blanket of a node *A* is composed by all of *A*'s parent nodes, child nodes and parent nodes of *A*'s child nodes. So from the learnt Bayesian network, we inferred the Markov blanket of the class label and found only six (6) moments {1, 6, 16, 21, 37 46} within the Markov blanket of the class label (see the shaded units in figure 5). These six (6) features represent the most optimal feature subset, because they have direct or indirect probabilistic dependence relationship with the class label.

Step 6: In order to determine the significance of the selected feature subset, we use it to train a Bayesian classifier and the 5-folds cross validation's classification accuracy was found to 83.82%, which is higher 3% than that based on all features in chi squared order and 6% higher than that based on all features in original order.

Table 2. Classification accuracy for feature subset selection

Feature Subset Size	Classifier	Accuracy
Phase I		
254 features	MLP	71.44%
254 features	SVM	79.41%
47 features in OMFS	MLP	74.00%
47 features in OMFS	SVM	86.96%
Phase II		
18 features in original order	Bayesian Network	77.21%
18 features in chi-square order	Bayesian Network	80.88%
6 Markov blanket features	Bayesian Network	83.82%

5. Concluding Remarks

From a practical standpoint, we demonstrated the potential of using Zernike moment methods as a viable image-processing approach for working with CSLT optic nerve images [16]. Furthermore, we presented a novel feature subset selection strategy that minimized the feature space without the loss of information. Table 2 indicates that through the first pass of our feature subset selection strategy we managed to reduce the feature set from 254 moments to a much smaller feature subset comprising just 47 salient moments, whilst achieving a slight increase in the classification accuracy. The second pass of our feature subset selection strategy, involves the use of a Markov Blanket as a filter model to the 47 features. We are able to further minimize the feature set to just 6 most salient moments whilst maintaining the classification accuracy.

We have presented an alternate approach to analyze CSLT images for glaucoma detection. In the next step we plan to derive symbolic rules using rule induction algorithms to provide symbolic knowledge for diagnosing glaucoma leading to the automation of decision support for glaucoma based on CSLT images.

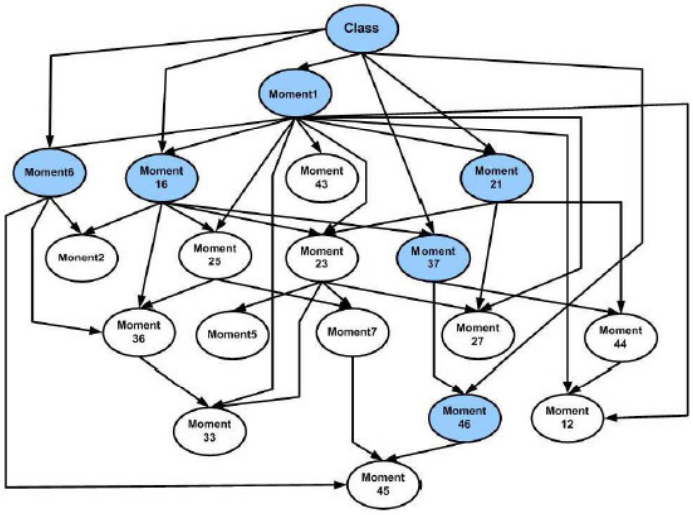

Figure 5. The learnt Bayesian network. The shaded nodes represent the Markov Blanket for the class label.

References

[1] G. Zinser, R.W. Wijnaendts-van-Resand, A.W. Dreher. "Confocal Laser Tomographic Scanning of the Eye", Proc. SPIE 1161, 1980, pp. 337–344,

[2] B.A. Ford, P.H. Artes, T.A. McCormick, M.T. Nicolela, R.P. LeBlanc, B.C. Chauhan. "Comparison of Data Analysis Tools for Detection of Glaucoma with The Heidelberg Retina Tomograph", Ophthalmology 110 (6) 2003, pp 1145-1150.

[3] Yu. V. Vorobyev, "Method of Moments in Applied Mathematics", Gordon and Breach Science Publishers, New York, 1965.

[4] C.H. Teh, R.T.Chin, "On Image Analysis by the Methods of Moments," IEEE Trans. Pattern Analysis by Machine Intelligence 10(4), July 1998, pp. 96-513.

[5] R. Kohavi, G. John, "Wrappers for Feature Subset Selection", Artificial Intelligence Journal , Vol. 97 (1-2).

[6] D. Koller, M. Sahami, "Toward Optimal Feature Selection", Proc. of the 13th Intl. Conf. on Machine Learning, 1996, pp. 284-292.

[7] I.Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning research, vol. 3, 2003, pp. 1157-1182.

[8] A.L. Blum, P.Langley, "Selection of Relevant Features and Examples in Machine Learning", Artificial Intelligence, 1997, pp. 245-271

[9] A.Khotanzad, J. Lu, "Classification of Invariant Image Representations Using a Neural Network", IEEE Trans Acoustics, Speech, and Signal Processing, Vol. 38, 1990, pp. 1028-1038.

[10] E.R. Hruschka Jr., E.R. Hruschka, N.F.F. Ebecken, "Feature Selection by Bayesian Network", Canadian AI Conference, LNAI 3060, Springer Verlag, 2004, pp. 370-279.

[11] C. Bowd, K. Cban, L.M. Zangwill, M.H. Goldbaum, T. Lee, T.J. Sejnowski, R.N. Weinreb, "Comparing Neural Networks and Linear Discriminant Functions for Glaucoma Detection using Confocal Scanning Laser Opthalmoscopy of the Optic Disc," Investigative Opthalmology & Visual Science, Vol. 43 (11), 2002.

[12] J. Park, J. Reed, Q. Zhou, "Active Feature Selection in Optic Disc Nerve Data using Support Vector Machine", IEEE World Congress on Computational Intelligence, 2002.

[13] N.V. Swindale, G. Stjepanovic, A. Cbin, F. Mikelberg, "Automated Analysis of Normal and glaucomatous optic nerve head topography images", Investigative Opthalmology & Visual Science, Vol. 41 (7), 2000.

[14] S. Cheng, Y. Huang, "A Novel Approach to Diagnose Diabetes Based on the Fractal Characteristics of Retinal Images," IEEE Trans. on Information Technology in Biomedicine, Vol.7 (3), 2003.

[15] A.Peters, B. Lausen, G. Michelson, O. Gefeller, "Diagnosis of Glaucoma by Indirect Classifiers", Methods of Information in Medicine, 2003.

[16] A.McIntyre, M. Heywood, P. Artes, S.S.R Abidi, "Toward Glaucoma Classification with Moment Methods", 1st Canadian Conference on Computer and Robot Vision (CRV04), 2004, London, Ontario.