An Automatic System for Classification of Nuclear Sclerosis from Slit-Lamp Photographs

Shaohua Fan[1], Charles R. Dyer [1], Larry Hubbard[2], Barbara Klein[2]
{shaohua, dyer}@cs.wisc.edu,
hubbard@rc.ophth.wisc.edu, kleinb@epi.ophth.wisc.edu
[1] Department of Computer Science, [2]Department of Ophthalmology & Visual Sciences
University of Wisconsin-Madison, USA

Abstract. A robust and automatic system has been developed to detect the visual axis and extract important feature landmarks from slit-lamp photographs, and objectively grade the severity of nuclear sclerosis based on the intensities of those landmarks. Using linear regression, we first select the features that play important roles in classification, and then fit a linear grading function. We evaluated the grading function using human grades as error bounds for "ground truth" grades, and compared the machine grades with the human grades. As expected, the automatic system significantly speeds up the process of grading, and grades computed are consistent and reproducible. Machine grading time for one image is less than 2 seconds on a Pentium III 996MHz machine while human grading takes about 2 minutes. Statistical results show that the predicted grades by the system are very reliable. For the testing set of 141 images, with correct grading defined by a tolerance of one grade level difference from the human grade, the automated system has a grading accuracy of 95.8% based on the AREDS grading scale.

1 Introduction

A cataract is a clouding or opacity of the eye's lens that can cause vision problems. Nuclear sclerosis is an important type of age-related cataract. Traditionally, the degree of nuclear sclerosis has been evaluated by a trained human grader based on comparison of the photograph to be graded with a series of standard photographs (called Standards). The grading system can use either an integer scale or a decimal scale. Grading systems using an integer scale include the Lens Opacities Classification System (LOCS) I-II system [3, 4], the Wisconsin system [7], an adaptation of which became the Age-Related Eye Disease Study (AREDS) system, the Wilmer system [12], the Cooperative Cataract Epidemiology Study Group (CCESG) system [9], and the Oxford system [10]. When a decimal scale is used, the grader places the photo between adjacent Standards, and then assigns a grade with a decimal value in the interval. Decimal scale cataract grading systems include the LOCS III system [5], Wisconsin AREDS lens grading protocol [1], and a simplified cataract grading system in the World Health Organization (WHO) cataract grading group [11].

While very useful, subjective systems have a number of disadvantages: (1) being a subjective process, the method shows large variability among graders (different people have

We would like to thank Dennis Hafford and Jane Armstrong for providing assistance and the slit-lamp photographs, Dr. Grace Wahba for discussions on statistical analysis methods, Dr. Nicola Ferrier for early collaborative discussions on feature detection methods, and Xin Qi for help in R programming. The support of the National Eye Institute of the National Institutes of Health under Grant Nos. N01-EY-0-2130 and EY-12652 is gratefully acknowledged.

different spectral sensitivity) and by the same grader over time; (2) the method is a manual process, so it is time-consuming; (3) as commonly utilized, the traditional method has limited capacity to account for the variability inherent in taking and developing photographs, such as exposure and development time; and (4) it is hard to reliably measure cataract severity change over time.

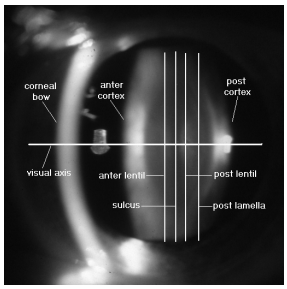
There have been some attempts towards computerized cataract detection recently [8]. However, so far there is no fully automatic and objective nuclear sclerosis grading system based on slit-lamp photographs. The goal of this work is to automate this process and provide an objective and repeatable grading system for nuclear sclerosis evaluation from slit-lamp images. Given a slit-lamp image of the eye, the system can automatically extract the feature landmarks in the image, and classify the level of nuclear sclerosis based on the intensities of those landmarks.

2 Materials and Photography Protocol

A Topcon SL-6E slit-lamp was used to take the nuclear sclerosis photographs. In this process, a vertical slit beam of light is shone through the lens nucleus at a 45° angle from visual axis after the pupil has been dilated pharmacologically, and the obliquely-illuminated lens is photographed with a camera situated on the visual axis. The result approximates a cross-section of the lens, depicting the backscatter of the beam as it travels through the lens nucleus from anterior to posterior. Slit lamp photographs are acquired as color slide transparences on Ektachrome 200 film. The slides were digitized on a Nikon CoolScan slider scanner. In our project, we use the AREDS grading system [1], which uses a decimal scale from 0.9 to 6.1.

There are about 1000 images used in this project. Those images are originally from the Beaver Dam Eye Study (BDES) and can be separated into four groups: 1) Standard Set includes the six base images, 1-6, from the AREDS grading system. These are the images to which other images are compared for grading. 2) Sample Set1 has 57 images that have been graded and the grades have been double-checked, so the grades for these images are more accurate and consistent. 3) Sample Set2 includes 93 images which were graded without double-checking. 4) Sample Set3 includes 800 images without human grades. In Sample Set1 and Sample Set2 there are a few poor-quality images that are rotated, blurred, or scaled. Those images were automatically detected and pulled out for human grading only. Sample Set1 has four images of this kind and Sample Set2 has five bad images.

3 Image Processing and Feature Detection



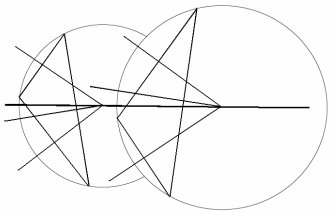


Fig. 1. Feature landmarks on slit-lamp image **Fig. 2.** Visual axis detection based on circle model

In our images, the corneal bow is the leftmost bright vertical curve in the image. Tracing left to right through the image, the corneal bow is followed by the dark anterior chamber. The leading edge of the anterior cortex is the second bright vertical curve. Other important features in the images are shown in Fig. 1.

3.1 Detecting the Visual Axis and Identifying the Ocular Landmark Features

The visual axis is the anterior-posterior line that bisects the nucleus horizontally. Since we will measure the degree of nuclear sclerosis based on the luminance values along this trace, reliable placement of this visual axis is critical for feature extraction and further analysis. The edges of the corneal bow and the anterior cortex are the most reliable features in the image, regardless of the degree of sclerosis, and hence were selected as the features used in identifying the location of the visual axis. In most cases, the corneal bow and the anterior cortex are symmetric with respect to the visual axis. If no noise existed, their edges could be modeled well as the arcs of two circles. Given an arc of a circle, we can calculate the center of the circle based on its curvature. After calculating the center points of the two circles defined by the corneal bow and the anterior cortex, the line connecting the two center points is a good approximation of the visual axis (Fig. 2).

We used the Canny edge detection algorithm [2] to detect edge points in an input slit-lamp image, then extracted the edges for the corneal bow and anterior cortex based on their relative locations in the image. However, explicitly fitting circles based on the edge data is problematic given the amount of image noise. One problem comes from the "keyhole" artifact in the anterior chamber, which is the reflection of the apparatus used. This artifact appears in almost every image, and its location in the anterior chamber varies from image to image. As a result, when we try to detect the anterior cortex edge, we may actually find edge points associated with the keyhole. We eliminated this problem by locating the keyhole using a matched filter technique. After the keyhole is detected, the edge points belonging to the keyhole were avoided when searching for the anterior cortex edge. The problem still exists in cases when the keyhole overlays the anterior cortex, however. The other problem comes from poor-quality images, which include blurred, rotated, scaled and improperly cropped images. Those problems can make the automatic process fail to detect the proper edges and therefore the estimated visual axis is poor.

In order to reliably detect the visual axis and automatically determine how well the visual axis is detected, we developed a voting scheme to obtain a confidence level indicating the correctness of the visual axis. This voting scheme essentially combines a Monte Carlo-like approach with a robust estimation technique [6].

The voting algorithm works as follows. Assume the edge of an arc is extracted and a circle model fits the arc reasonably well. Our method to estimate the center of the circle is: (1) randomly select five widely-separated points on the arc. From these five points, identify the five longest chords between pairs of these points; (2) compute the lines perpendicular to the five chords; (3) find the intersections of these lines. If the arc is perfectly circular, all five lines will intersect at one point; otherwise, there will be 10 intersections; and (4) compute the centroid of the 10 intersections as the location of the estimated center of the circle.

The circle center for the corneal bow is found in two rounds. In the first round, repeating the above four steps many times with different sets of five edge points, each time an approximate center is obtained for the circle. By averaging these center point estimates, we get the center, C1, of the circle based on all selected sets of five points. Note in the first round, there may exist some outliers in the circle center approximation. In the second round, we repeat the steps used in first round, except we use the approximated center C1

to eliminate outliers, i.e., centers whose distance from C1 is larger than a threshold. We average all the center points except the outliers in the second round and denote the centroid of these points C2. This two round process produces a robust estimate of the center even when there is a high percentage of outlier edge points used.

The same algorithm is applied to estimate the center of the circle for the anterior cortex. At each iteration in the second round, after we get an approximate center of the circle for the anterior cortex, we connect C2 with this center to form a visual axis line. Since the line always goes through the fixed point C2, we can use a one-dimensional array to record the orientation of these lines. After many iterations, the histogram of the orientations provides evidence of the best location of the line (with highest consensus) and the distribution of orientations gives a measure of the uncertainty (Fig. 3).

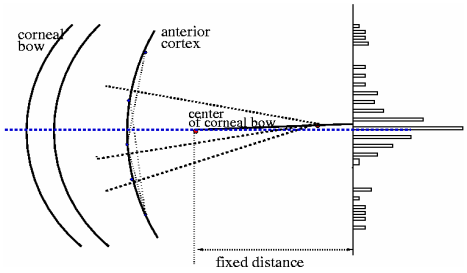


Fig. 3. Voting scheme for visual axis detection

In summary, the following algorithm was used to automatically detect the visual axis in a slit-lamp image:

Step 1. Monte Carlo Simulation

Randomly choose a set of five widely-separated edge points and calculate the center of the circle estimated using these five points.

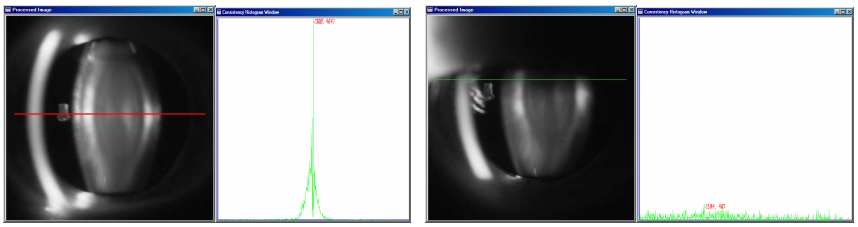
Step 2. Voting for Visual Axis Detection

i) Repeat Step 1 many times for corneal bow edge points in two rounds. Throw out center point outliers in the second round and average the remaining center points to get a center point of the corneal bow. Use this center point, C2, as one point that defines the visual axis.

ii) Repeat Step 1 many times for anterior cortex edge points in two rounds. Throw out center point outliers in the second round. At each iteration in the second round, compute the line connecting C2 and the calculated center point. Histogram the orientations of all the estimated visual axis lines.

iii) Select the line orientation that occurs with the highest frequency in the histogram as the visual axis.

This approach is efficient and robust to image noise. At each iteration, only five points are used, so computing the center for the best circle fit by those five points is fast. However, using a true Monte Carlo approach requires some criterion to evaluate the results of each iteration. Without any such measure available, accumulating all estimates and taking the one with the "most votes" maintains the benefits of randomization: statistically, we are likely to obtain valid solutions most of the time. The iterations enable us to eliminate image noise as arcs fit to erroneous data produce centers distributed over a large area. Arcs along the true circle "agree." Hence we can determine, by consensus, the best center, and by looking at the distribution of centers, we can determine a level of uncertainty (Fig. 4).



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

Fig. 4. Visual axis orientation voting histograms for two images. (b) Line orientation histogram for image (a), (d) Line orientation histogram for image (c). The single sharp peak in (b) indicates that most iterations agreed with the line location. A broader histogram in (d) indicates a lack of consensus, and therefore some difficulty in ascertaining the correct line location

After the visual axis is detected, a linear densitometric track along the visual axis was performed to get a one-dimensional intensity profile. To make the values of the profile robust, a narrow band centered on the visual axis was used to average the pixels along the vertical direction in the band at each point on the visual axis. Based on this intensity profile, the original image and the edge image, landmark features can now be detected reliably with conventional image processing methods.

3.2 Construct Regression Features

Seven landmark features were extracted from the photograph, and, for each landmark, an intensity mean of the pixels within a 7 pixel wide by 60 pixel high rectangle centered on this point was recorded. Based on expert knowledge, four of the seven features were chosen as most significant: anterior lentil, sulcus, posterior lentil, and posterior lamella (with variable names AnteLen, Sulcus, PostLen, and PostLam, respectively). The other three features are highly correlated with the four selected ones and thus considered to have negligible effect on grading. During the human grading procedure, the trend of intensity change from anterior lentil to posterior lamella along the visual axis plays an important role, so several composite features were also defined, for example, the ratio between the intensity at the anterior lentil and the posterior lentil (RatioALPL, RatioALPLNorm).

Two other features considered were the standard deviation of intensity in the neighborhood of the sulcus in order to take into account the effect of intensity in the center of the lens. An eye whose image has a narrow black strip along the sulcus is defined to have low degree of nuclear sclerosis. Two areas of interest of different size were defined: one is a 7 pixel by 60 pixel rectangle, and the other is 17 by 120, both centered on the sulcus. The feature *SulStdSmall* is the intensity standard deviation of the pixels in the small rectangle, and the feature *SulStdLarge* is the intensity standard deviation of the pixels in the big rectangle. Intuitively, the ratio of *SulStdSmall* to *SulStdLarge*, which defines the feature *ratioSulStd*, should be small if a narrow black strip exists around the sulcus. In summary, 10 features were computed from each image and used to represent all the information for grading.

4 Data Analysis and Results

We built our grading function in two steps. First, the Standard Set and Sample Set1 were used as the training data to select the most important features for grading. Second, the Standard Set alone was used to train the grading function, which was defined using the

most important features selected in the first step. Strictly speaking, the only "ground truth" data are the Standards because the grades assigned to the Standards are the definition of the grades. However, six images are too few to estimate the parameters in a grading function with many variables. Since the grades for the images in Sample Set1 were checked after grading, they are more accurate and consistent. Even though they can not be treated as ground truth due to the limitations of human grading, those scores provided a good approximation of the "correct grades," so it is reasonable to use those data to determine the most important features affecting grading. After the most important features were identified, the Standards were used to determine the parameters for combining the features. We tested the grading function using Sample Set1 and Sample Set2 as testing data.

4.1 Feature Selection

To decide which of 10 candidate features are important for grading, a linear regression model using all 10 features was used to fit the training data (Standard Set and Sample Set1). Table 1 shows the coefficients of the model and the *p*-values for the 10 features and the constant term.

Table 1. Coefficients for a linear model with 10 features[1]

	Estimate	t value	Pr(>|t|)
AnteLen	-0.2704	-3.134	0.0033 **
Sulcus	0.1012	5.678	1.45e-06 ***
SulStdSmall	0.1227	0.256	0.7990
SulStdLarge	-0.1343	-0.502	0.6181
PostLen	0.2163	2.760	0.0088 **
PostLam	-0.0030	-0.211	0.8342
RatioALPL	-1.5422	-0.916	0.3651
RatioALPLNorm	36.4366	3.722	0.0006 ***
RatioALPostLam	0.3806	0.760	0.4519
RatioSulStd	-0.1483	-0.104	0.9179
(Intercept)	-34.7648	-3.606	0.0009 ***

The features with *p*-values less than 0.001 are considered significantly important for grading. Those features are *Sulcus* and *RatioALPLNorm*. This is consistent with experts' knowledge. The intensity of the sulcus is a good indicator of the general brightness of the image, and the images of eyes with more serious nuclear sclerosis tend to be brighter. The feature *RatioALPLNorm* indicates the intensity change from anterior lentil to posterior lentil. The higher the *RatioALPLNorm* value, the higher the degree of nuclear sclerosis the eye usually has. The normalized ratio, *RatioALPLNorm*, is less sensitive to image noise due to variations of exposure time or development time compared to the un-normalized value, *RatioALPL*; this improvement is evident from the fact that *RatioALPLNorm* has a lower p-value than *RatioALPL*. The features designed to detect the strip around the sulcus, *SulStdSmall*, *SulStdLarge* and *RatioSulStd,* did not have low p-values, probably because of the difficulty of identifying this information.

4.2 Model Fitting and Accuracy Evaluation

[1] Significant level codes for *p* value: "***": [0,0.001); "**":[0.001, 0.01); "*":[0.01,0.05); ".":[0.05,0.1); " ":[0.1,1]

After the two most important features for the classifier were determined, a linear grading function using those features was fit using the Standards. To evaluate how well the grading function predicts the grades for the testing data, the computed decimal grades were converted to the scale used in the AREDS system by rounding the grade to the nearest 0.1. The difference between human grades and computed grades was quantized into 40 groups (-5 to 5 at intervals of 0.25). The histogram of the grading differences is a good indicator of how close the machine grading is to the human grading. The grading function as a linear combination of the features *Sulcus* and *RatioALPLNorm*, was defined by:

$$\text{Grade} = 0.03077*\text{Sulcus} + 1.40517*\text{RatioALPLNorm} - 0.4654$$

Table 2 shows that the computed grades are very close to the real values for the Standards.

Table 2. Computed grades for the Standards

Standard	1	2	3	4	5	6
Computed Grade	0.98	1.96	3.14	3.98	4.91	6.03

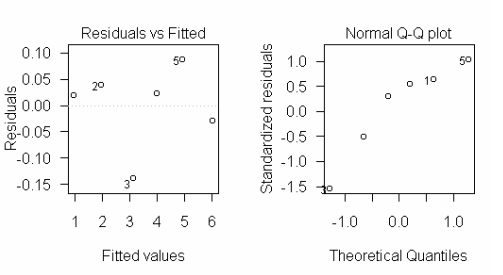


Fig. 5. *Residuals vs Fitted* plot and *Q-Q* plot

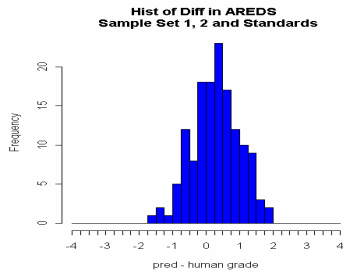


Fig. 6. Histogram of grading differences between human grades and computed grades

The normal Q-Q plot in Fig. 5 shows the residuals are very close to being normally distributed. The plot of the residuals and their corresponding grades show there are no trends or patterns in the residuals, visually verifying the use of a linear model.

For the grades of the testing images predicted by the grading function, Fig. 6 shows that out of 141 images, 135 are machine graded to within one grade of the human grade, which is 95.8% of the population. No image has more than a two grade difference, and only 6 images have a two grade difference. In human grading, one grade fluctuation is quite common and regarded as acceptable.

5 Concluding Remarks

A system has been developed that automatically detects the visual axis and extracts features from slit-lamp photographs. Expert knowledge and a linear regression model were used to define and select important features for nuclear sclerosis grading. After the two most important features were chosen, a linear grading function was fit using the Standards and evaluated based on human grading. The linear grading function achieved a grading accuracy of 95.8% within 1 grade using the AREDS grading system for the testing data. While adding some relatively important features such as anterior lentil and posterior lentil into the grading function may slightly reduce the residual standard error for prediction in some test cases, if similar classification accuracy is achieved, the grading function with

fewer features is preferred since it has the advantage of being simpler and more robust to image noise and image processing errors.

It is interesting to point out that we achieved this classification result using only the six Standards as training data. This is evidence that the linear grading function using the two features, *Sulcus* and *RatioALPLNorm*, largely captures the relationship between the severity of nuclear sclerosis and the image. With a function of known analytical form, the two unknown parameters can be determined by two good samples. This may explain why using only six Standards results in a good grading function.

To further evaluate classification accuracy, human graders can look at the machine grades to see how many of them are acceptable. The next step is to automatically grade the 800 images in Sample Set3 and ask human experts to evaluate them. Other images to be tested include the Standards under different exposure or development times, and follow-up images for individuals over time. Since selecting important features is critical for modeling the grading function, it may be worth trying some more complex feature selection methods such as Likelihood Basis Pursuit (LBP) [13]. Another possibility is to consider non-linear grading functions.

References

1. The Age-Related Eye Disease Research Group: The Age-Related Eye Disease Study (AREDS) System for Classifying Cataracts from Photographs. AREDS Report No. 4. *Am J Ophthalmol.* 131 (2001) 167-175.
2. Canny, J.F.: A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 8, No. 6 (1986) 34-43.
3. Chylack, L.T. Jr, Leske, M.C., Sperduto, R., Khu P., McCarthy, D.: Lens Opacities Classification System. *Arch. Ophthalmology*. 106 (1988) 330-334.
4. Chylack, L,T. Jr, Leske, M.C., McCarthy, D., Khu, P., Kashiwagi, T., Sperduto, R.: Lens Opacities Classification System II. *Arch. Ophthalmology*. 107 (1989) 991-997.
5. Chylack, L.T. Jr, Wolfe, J., Singer, D., McCarthy, D., Carmen, J., Rosner, B.: Quantitating Cataract and Nuclear Brunescence, The Harvard and LOCS System. *Optometry and Vision Science*. Vol. 70, No. 11 (1993) 886-895.
6. Fischler, M., Bolles, R.: Random Sample Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography. *Comm. ACM.* 24 (1981) 381-395.
7. Klein, B., Klein, R., Linton, K.L.P., Magli, Y.L., Neider, M.W.: Assessment of Cataracts from Photography in the Beaver Dam Eye Study. *Ophthalmology*. 97 (1990) 1428-1433.
8. Robman, L.D., McCarty, C.A., Garrett, S.K., Stephenson, H., Thomas, A.P., McNeil, J.J., Taylor, H.R.: Comparison of Clinical and Digital Assessment of Nuclear Optical Density. *Ophthalmic Research.* 31 (1999) 119-26.
9. Sasaki, K., Sakamoto, Y., Fujisawa, K., Kojima, M., Shibata, T.: A New Grading System for Nuclear Cataracts - An Alternative to the Japanese Cooperative Cataract Epidemiology Study Group's Grading System. *Cataract Epidemiology*. 27 (1997) 42-49.
10. Sparrow, J. M., Brown, N.A.P., Shun-Shin G. A. and Bron A. J.: The Oxford modular cataract image analysis system. *Eye*. 4 (1990) 638-648.
11. Thylefors B., Chylack, L.T. Jr, Konyama, K., Sasaki, K., Sperduto, R., Taylor, H.R., West, S.: A Simplified Cataract Grading System. *Ophthalmic Epidemiology*. 9 (2002) 83-95.
12. West, S.K., Rosenthal, F., Newland, H.S., Taylor, H.R.: Use of Photographic Techniques to Grade Nuclear Cataracts. *Invest Ophthalmol Vis Sci.* 29 (1988) 73-77.
13. Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R., Klein, B.: Variable Selection and Model Building via Likelihood Basis Pursuit. TR# 1059, Dept. of Statistics, UW-Madison, (2002).