Measurement of Bulbar Redness

Paul Fieguth, Department of Systems Design Engineering, University of Waterloo
Trefford Simpson, School of Optometry, University of Waterloo

Keywords: Contact lenses, Image Analysis, Statistics, Clinical Grading, Hyperaemia, Pattern Recognition, Clinical Scales, Redness

Abstract

Purpose: To examine the relationship between physical image characteristics and the clinical grading of images of conjunctival redness; to develop an accurate and efficient predictor of clinical redness from the measurements of these images.

Methods: Seventy-two clinicians graded the appearance of 30 images of redness on a 100 point sliding scale with three referent images (at 25, 50 and 75) through a WWW-based survey. Using software developed in MATLAB, each image was quantified in two ways: (i) the presence of blood vessel edges, based on the Canny edge detection algorithm, and (ii) a measure of overall redness, quantified by the relative magnitude of the redness component of each RGB pixel. Linear and nonlinear regressors and a Bayesian estimator were used to optimally combine the image characteristics in order to predict the clinical grades.

Results: The clinical judgments of the redness images were highly variable: The average grade range for each image was about 55, more than half the extent of the entire scale. The median clinical grade was chosen as the most reliable measure of “truth”. The median grade was predicted by a weighted linear combination of the edgeness and redness features of each image. The strength of the predicted association was $\tau = 0.976$, exceeding the strength of association of all but one of the 72 individual clinicians.

Conclusions: Clinical grading of redness images is highly variable. Despite this human variability, easily-implemented image analysis and statistical procedures were able to reliably predict median clinical grades of conjunctival redness.

1 Introduction

The clinical judgment of ocular redness is complex and poorly understood. Typically, the appearance of the eye is judged based on a scale, and the examination of these scales provides a lesson in contemporary views of measurement. Even the simplest binary descriptive scale (red and not red) may be regarded as quantitative with the data provided being either nominal or ordinal\(^1\). Other classifications include those based on the underlying reference of the scale (verbal or visual) and the numerical basis of the scale, whether discrete\(^2\) or continuous\(^3\).

Theoretical examination aside, the scales themselves are typically poorly described and with few exceptions have been untested\(^2\). In addition to a lack of understanding of the scales themselves, there is no empirical information about how clinicians make judgments of redness. Indeed, our data show evidence to suggest that clinicians quote wildly inconsistent grades, even in the presence of a well-defined grading scheme. Figure 1 summarises the motivation of this paper: it was arranged\(^4\) for 72 subjects to grade the clinical appearance of the redness of thirty different pictures of conjunctivas. The figure shows the results arranged in order of ascending median redness (solid line), and plots the quartile ranges. As is apparent, for each image the range of redness estimated by the subjects was at least 25% of the total scale, and on average in excess of 55% of the total scale. These results clearly show the extremely poor quantitative accuracy of such clinical grading, and the degree of subjectivity which is present in human grades. In light of Figure 1, this paper presents an automated, objective alternative.
Clinical grading may be judged using at least two general strategies. The first is primarily luminance / chromaticity based. Judgments are made on the basis of the overall redness and brightness (luminance) of the eye. As the redness goes up, so the luminance goes down. A second strategy is made on the basis of the appearance of the visible vessels. This could include judgments of the diameter of vessels, vessel tortuosity, and the proportion or number of vessels occupying the area to be graded. The difference between these methods is really one of scale: Luminance judgments would correlate with vessel-appearance if the capillary beds giving rise to the conjunctival flush were resolved. Similarly, with a sufficiently low resolution, smaller vessels would not be resolved and would "blend" into the background redness. For any typical clinical observation, however, each type of judgment is possible and could vary (to a large extent) independently of the other. The automated / objective approach proposed in this paper will be based on the same two criteria; two features will be extracted from each image, one based on redness, and the other based on the appearance of blood vessels.

Because of the vagaries of clinical scales and grading there have been a number of past attempts to perform clinical grading using automated methods. These have typically involved examining the structure in a particular area to determine the characteristics of the vessels. Most recently, Papas showed that the clinical grade of redness of a relatively small patch of conjunctiva was strongly linearly associated with an automated technique that measured the number of vessels in the patch. This very interesting result is unfortunately difficult to relate to clinical grading, however, since the regions evaluated were relatively small and the task was somewhat different from typical clinical grading where, usually, almost the whole nasal and temporal bulbar area is graded.

This paper proposes to study the relationship between clinical grading and quantitative aspects of conjunctiva images, with the goal of developing an automated estimator for conjunctival hyperaemia. The purpose of the estimator is to reproduce the overall trend, but to eliminate the inconsistent and irreproducible details, of the clinical ratings. Quantifiable features were correlated with the clinical grading data to produce an estimator that is accurate, consistent, and repeatable.

2 Methods

Data Collection

Thirty images of bulbar redness were used. These ranged in redness from normal to severe. The images were derived from frontal photographs taken with constant magnification and diffuse white illumination and included enough of the lateral and medial canthi to recognize nasal and temporal bulbar conjunctiva. The images rated as the least and the most red are shown in Figure 2.

A web-site was developed to display the images and to collect the ratings. Subjects were required to grade each image, presented in a random order, on a zero to 100 point scale using sliders for both the nasal
and temporal bulbar areas. Because of the inconsistencies between the computer monitors of different subjects, for example the brightness, contrast, and gamma settings, each page of the survey included three smaller images that in a previous experiment were shown to represent approximately the levels 25, 50 and 75 on a 100 point scale. Figure 3 shows a typical display/rating page. The whole survey could be completed in approximately ten minutes, so user tedium should not significantly impact the quality of the collected ratings.

Image Analysis

The least obvious step in our analysis is the determination of quantitative, mathematical aspects of an image of the eye which correlate with the grades as assessed by clinicians. The survey on the website did permit respondents to describe the criteria by which they passed judgment, however even relatively precise statements such as “average artery width” or “average redness” are not readily represented as an image processing algorithm because of the vast number of subjective and subconscious operations undertaken by the human visual system.

Instead, we propose to analyze redness on the basis of two straightforward features, based on a model of the trauma mechanism. Conjunctival hyperaemia is characterised by the expansion of small arteries just below the surface of the eye. As the blood vessels swell they become much larger and easily detected as a red line on a white scleral background — we propose to use an edge-detector (specifically the Canny method to measure the total length of visible arteries. However the smallest arteries are resolvable neither by the pixels in a CCD camera, nor by the human eye, so a mild onset of hyperaemia begins as a diffuse reddening with no discernible edges — for such cases we propose an integrated measure of redness.

Figure 2: The best (a) and worst (b) samples of our thirty-image set.

Figure 3: A screen-shot from the survey web-site. The three benchmark images are always visible to the user at the bottom of the screen. The two grades (temporal and nasal) are set using the sliders on the right.
We do not maintain that these represent the “optimum” features, rather the rationale is that if the performance using just these simplistic methods is good, then clearly additional study and criterion refinement can only lead to a further improvement in the results.

Each image $I$, with composite components $I_R, I_G, I_B$, is segmented into two nasal and temporal subsets, $S_R, S_B$, respectively, to allow the two sides to be analyzed separately. The redness feature

$$f_r(S) = \frac{\sum_{i \in S} [2[S_R]_i - [S_G]_i - [S_B]_i]}{2 ([S_R]_i + [S_G]_i + [S_B]_i)}$$

represents the average integrated redness in subimage $S$. Note that black pixels, which have no defined colour, have been removed from $S$. The denominator term normalises the feature, such that $-0.5 < f_r < 1$:

Red Image $\iff f_r = 1$
White, Grey Image $\iff f_r = 0$
Blue, Green Image $\iff f_r < 0$

Even the most seriously traumatised eye is not completely red; the feature range for the images in our experiment was approximately $0 \leq f_r \leq 0.25$.

The edge feature

$$f_e(S) = \frac{\sum_{i \in S} \text{Canny}(S)_i}{[S]}$$

returns the fraction of pixels which are identified as edges; that is, the ratio of the number of edge pixels, computed by a Canny edge detector, to the total number of pixels $[S]$. The premise of the edgeness feature is that perceived redness is not just a function of average colour, but also the number or density of arteries.

In preparing the segmented subsets $S_R, S_B$, care had to be taken to eliminate skin pixels, whose reddish hue would bias redness feature results, surrounding hair, whose strong contrast would affect the edge feature, the pupil and the iris. To ensure the accuracy of the results this segmentation was carefully performed by hand, although automating this step should be straightforward, because the colour of the sclera is quite distinct from its surroundings.

Figure 4: The median-removed distribution of grades. The distribution is approximately Gaussian, with an odd periodicity due to the human bias towards rounded numbers (multiples of five and ten).

3 Results

We have two sets of data: The grades collected from clinicians through the survey, and the computed values of the redness and edge features.

Grading Analysis

Each of the thirty different eye images were graded by a total of seventy-two different clinicians. Although the results show a broad degree of consensus, there is an astonishing variability from person to person, shown in Figure 1. The average range in the grades is 55, more than half of the entire range of the scale. Even the three calibration images are no exception: Although users were told to grade the middle calibration image as 50, the grades assigned to that image had a tremendous range from 22 to 90. This range is not due to the tedium of completing the survey, since the variability was not observed to increase with the position (early versus late) in the survey.

The histogram of the distribution of assigned grades about the median is shown in Figure 4. The distribu-
tion is vaguely Gaussian, with an odd superposition of spikes. Further analysis of the data shows that the spikes are due to the human preference for rounded numbers: Multiples of five are more than four times as prevalent as other numbers, lending further support to the need for an objective grade.

The standard deviation of the grading distribution for each eye varies between approximately 6% and 15%. The trimmed standard deviation, based on keeping only the 50% most consistent grades, is considerably tighter (as implied by Figure 1) at 2.9% to 8%.

**Feature Regression**

Figures 5 and 6 show the raw data points of the redness and edgeness features, respectively, versus the median human grade. There is a clear relationship between the features and the human data, although the relationship is not necessarily linear (especially for redness feature $f_r$), and may have varying degrees of consistency (for example the three or four outliers in the edgeness data in Figure 6).

Our goal is to predict, in some fashion, the grade from the extracted image data. We will denote by $\tilde{g}(f)$ the estimated grade $g$ based on feature value $f$. Clearly we want to constrain the grade to lie within the scale:

$$0 \leq \tilde{g}(f) \leq 100.$$

The solid lines in the figures represent the chosen regressions. Because of the wide range of $f_r$ (up to 1.0), a linear fit is inappropriate, so a hyperbolic regression was chosen for $f_r$, having an asymptote at $g = 100$ and a slope at the origin of 45/0.05. Although unenlightening, for completeness the temporal-redness regression follows:

$$\tilde{g}^2(f_r) - 900\tilde{g}(f_r) \cdot f_r - 109\tilde{g}(f_r) + 90000f_r - 270 = 0$$

Although this may appear over-fit, the equation was fit by adjusting only one free parameter, once the slope and asymptote were specified.

A more straightforward linear regression was chosen for $f_e$, where the three misfitting data points were eliminated from the coefficient learning process. The resulting expression for the nasal-edgeness regres-
Sion is
\[ \hat{g}(f_e) = 5 + 60 \cdot f_e / 0.16 \]  
(6)

With two estimators \( \hat{g}(f_r), \hat{g}(f_e) \) defined there is clearly an ambiguity regarding which estimator to use, or whether the estimators can somehow be combined automatically. If we view \( \hat{g}(f_r), \hat{g}(f_e) \) as approximate “measurements” of the true grade \( g \), then under certain conditions\(^1\) the optimal linear Bayesian estimate of the grade is
\[ \hat{g}(f_r, f_e) = \frac{\hat{g}(f_r) \sigma_r^2 + \hat{g}(f_e) \sigma_e^2}{\sigma_r^2 + \sigma_e^2} \]  
(7)

and the associated estimation error variance is
\[ \text{var}[\hat{g}(f_r, f_e)] = \frac{1}{\sigma_r^2 + \sigma_e^2} \]  
(8)

where \( \sigma_r^2, \sigma_e^2 \) are the error variances of the single-feature estimators \( \hat{g}(f_r), \hat{g}(f_e) \) respectively. These error variances cannot be deduced theoretically, but need to be inferred from the data. We have computed them as the smoothed local sample variance of the human grades about the regressed curves; the resulting one-standard-deviation curves are shown in Figures 5 and 6. Clearly the Bayesian estimator (7) will bias in favour of estimator \( \hat{g}(f_r) \) for eyes having only mild redness, and towards \( \hat{g}(f_e) \) for severe eyes.

The above developments were discussed and illustrated, for compactness, based on only one half of the data, ignoring the nasal-redness and temporal-redness cases. In the following results all of the data are used.

Figure 7 shows the estimation results, using the Bayesian estimator (7) for both temporal and nasal data. The estimation results lie very close to the dashed ideal, with a correlation coefficient of 0.976 between the estimates and the human medians. For comparison purposes, an equivalent plot is shown in Figure 8, where a statistical sample of the human grades is plotted against the median, for a corresponding correlation coefficient of only 0.841.

The error bars in Figure 7 are unit standard-deviation in length, based on the Bayesian error variance (8). If the error variances are accurate, they should meaningfully reflect the distribution of the estimates \( \hat{g} \)

\(^1\) Ideally, \( \hat{g}(f_r), \hat{g}(f_e) \) need to be unbiased estimates of grade \( g \), and the errors in the two estimates are assumed to be independent.
Figure 9: Standard deviations of the estimation errors: + clinicians, o the 50% most consistent clinicians, and • the proposed automated system. Except for grades at the severe end of the scale, where the data were sparse, the errors of our proposed system are as good as or better than the 50% consensus of a group of 72 graders.

Figure 10: The performance of individual clinicians relative to the proposed system: of 72 respondents, all but one performed more poorly than our feature-based automated system. The dashed line shows the correlation of 0.976 attained by the method of this paper.

Clearly our errors are competitive with or better than even the most consistent graders.

4 Discussion

The results of this study reinforce previous results: Automated measures indeed provide information that is linearly associated with subjective grades of redness. Our results are similar to those of Willingham et al. and Papas in that we each found strong associations between the subjective grades and the measurements, as opposed to the weaker associations of Guillon and Shah. Our methodology is more similar to that of Willingham et al and Papas inasmuch as we used images that were graded, while Guillon and Shah’s subjective data were collected “in vivo” from a slit lamp biomicroscope. Our methods differ from previous workers who have either not used first order (overall redness) information or have used it separately from second order (vessel attribute) information. We provide a novel, straightforward method for the combination of image fea-
tasures that is remarkably concordant with the grades assigned by clinicians.

A primary objective of this study was to minimise the required operator intervention. Some previous studies have utilised combinations of custom software/hardware, making the analysis inaccessible or expensive to develop, operate, and maintain. For our research, "normal" Pentium PC's running Windows performed the data acquisition, and numeric processing was done using the MATLAB environment (widely available). Operator intervention was limited to a few mouse clicks to assist in the segmentation of the eye, removing the lids and the corneal components of the images. These processing steps can therefore can be implemented almost universally.

The way we chose to obtain grading was somewhat unusual, and perhaps controversial, in that we were unable to control our subjects and that our sampling method was far from randomised. These experimental attributes are no different, however, from any of the previous reports comparing automated methods to subjective methods of grading. Our sampling method did provide additional diversity, in that subjects were not from a single institution. The associated diversity in skill set also provided a more realistic sense to the grading data, in that not all subjects were true experts utilising grading scales many times per day. In other words, despite these additional sources of variability the clinical data were still remarkably well predicted by the proposed automated measures.

The Introduction stated that very little is known about grading techniques. Of particular importance in this regard is Figure 4, showing clear peaks at decimal and mid decimal values, similar to the effects observed in the literature. This was not accomplished by accident, because the subjects would have needed to carefully adjust a slider to generate these numbers. This suggests strongly that there is a tendency not to use the many steps on a 100-point scale and that, perhaps, all that is required is a 20 point redness scale. There are theoretical and practical implications of this, but the exact impact on the accuracy and repeatability of redness grading would need to be determined empirically.

Another result relating to grading was the large range of grading associated with the reference images which were part of the data set that clinicians graded. Although the median grades were very similar to the reference grades (Figure 1), there was a surprisingly large range of grades associated with each reference. This suggests that subjects either cannot psychophysically match grades to the references, an unlikely conclusion, or that subjects chose to "ignore" the values assigned to the reference. Clinicians have been shown to resist using tools which are assistive and this is perhaps a manifestation of this phenomenon. The clinician "disagrees" with the grade assigned to the reference and simply ignores it.

The results show that there are strong associations between the computed and clinically assigned grades, Figures 5 and 6 illustrate the distribution and variability of the individual computed grades. The error bars on Figure 7 show the variability of the combined computed grade for each image. In comparison, estimates of the variability of the clinical graders' performance is illustrated in Figure 8 using resampling techniques. The point illustrated in comparing the latter two figures is that there is more precision in the estimates using the computed grades, clearly illustrated in Figure 9, which compares the standard deviation of the computed approach for each slide with the actual standard deviation of the graders.

The experiment was developed from questions of grading, however the data may provide information pertinent to other areas. For example, how should images be compressed or coded for teledicine applications? Image compression reduces storage or transmission needs, but may also be associated with a loss of information. All of the images in our analysis (and web survey) were stored in a lossless (tiff) form precisely because it is unknown which attributes of eye images may be discarded without removing critical information. By better understanding which information is needed clinically, more effective com-
pression may be developed that minimises the loss of critical information at the expense of unimportant image content, for example by examining the changes in the image features \( f_r, f_c \) as a function of the compression type. This is similar to previous suggestions using perception to constrain image coding, except that the data are clinically salient rather than only perceptually salient.

In conclusion we have shown that computational techniques may be used to measure the redness characteristics of images of the bulbar conjunctival areas of eyes. These estimates compare very well to those derived using clinical grading methods. In addition the method we propose is much less variable than variability which exists between clinical graders. In the last 10 years numerous methods have been proposed to assign redness scores using computational methods. One might then ask, how does the procedure developed here advance these methods? For example, Villumsen et al. had strong correlations between a computational and a grading redness estimate. We believe that there are a number of reasons why this experiment describes methods and results that actually make it feasible to use this technique as a replacement for grading the redness of images. First, the technology is readily available and inexpensive. Whereas some previous studies have used rather exotic hardware/software combinations, the algorithms we used are available to anyone using a computer running almost every operating system that will be encountered on a desktop. Second, a minimum amount of operator intervention is required. This removes some of the subjectivity in some previous techniques and further lends itself to automation. And finally, we have shown that both accuracy and much less variability are present in the automated technique than in the subjective technique. These factors, we believe, provide a strong rationale for the adoption of this technique to replace clinical grading of bulbar redness. Because anterior segment assessment is much more than just redness evaluation, how to implement this technique more generally to replace in vivo grading is yet to be determined.

Acknowledgements

We would like to thank all of the observers who visited the web site and completed the grading task. We would also like to thank the following former students for their dedicated efforts in preparing the survey website and collecting the grading data: J. Cullen, C. Handler, V. Nagaraj, S. Nichols, S. Pounder, and K. Whitear. Finally our appreciation to the anonymous reviewers who provided helpful suggestions in improving the manuscript.

References


