

# Automated Measurement of Bulbar Redness

Paul Fieguth<sup>1</sup> and Trefford Simpson<sup>2</sup>

**PURPOSE.** To examine the relationship between physical image characteristics and the clinical grading of images of conjunctival redness and 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 points) through a World Wide Web-based survey. Using software developed in a commercial computer program, each image was quantified in two ways: by the presence of blood vessel edges, based on the Canny edge-detection algorithm, and by a measure of overall redness, quantified by the relative magnitude of the redness component of each red-green-blue (RGB) pixel. Linear and nonlinear regressors and a Bayesian estimator were used to optimally combine the image characteristics to predict the clinical grades.

**RESULTS.** The clinical judgments of the redness images were highly variable: The average grade range for each image was approximately 55 points, 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  $r = 0.976$ , 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. (*Invest Ophthalmol Vis Sci.* 2002;43:340-347)

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.<sup>1</sup> Other classifications include those based on the underlying reference of the scale (verbal or visual) and the numerical basis of the scale, whether discrete<sup>2</sup> or continuous.<sup>3</sup>

Theoretical examination aside, the scales themselves are typically poorly described and with few exceptions have been untested.<sup>2,3</sup> In addition to a lack of understanding of the scales

themselves, there is no empiric 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 summarizes the motivation of this article: Arrangements<sup>4</sup> were made for 72 clinicians to grade the clinical appearance of the redness of 30 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 graders 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 that is present in human grades. In light of the data in Figure 1, this article 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 increases, so the luminance decreases. 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 and objective approach proposed in this article is based on the same two criteria: Two features are 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 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.<sup>5-8</sup> Most recently, Papas<sup>9</sup> 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, because 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.<sup>2,3,10</sup>

This was a study of the relationship between clinical grading and quantitative aspects of conjunctiva images, with the goal of developing an automated estimator for conjunctival hyperemia. 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.

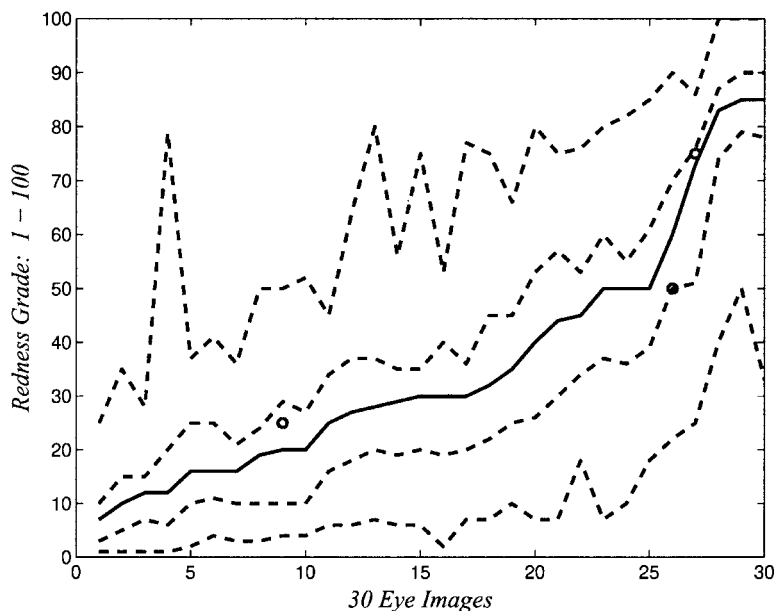
From the <sup>1</sup>Department of Systems Design Engineering and the <sup>2</sup>School of Optometry, University of Waterloo, Waterloo, Ontario, Canada.

Submitted for publication April 4, 2001; revised October 1, 2001; accepted October 18, 2001.

Commercial relationships policy: N.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Corresponding author: Paul Fieguth, Department of Systems Design, University of Waterloo, Waterloo, Ontario, Canada, N2L-3G1; pfieguth@uwaterloo.ca.



**FIGURE 1.** The extent of variations in human grading. *Bottom to top:* the five curves indicate the minimum, 25th percentile, median, 75th percentile, and maximum grade for each of 30 images. The scale was defined using three benchmark images, at grades of 25, 50, and 75 points, indicated by circles.

**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<sup>4</sup> to display the images and to collect the ratings. The observers were required to grade each image, presented in a random order, on a 0- to 100-point scale, using sliders for both the nasal and temporal bulbar areas. Because of the inconsistencies between the computer monitors of different graders, for example the brightness, contrast, and gamma settings, each page of the survey included three smaller images that in a previous experiment<sup>3</sup> were shown to represent approximately the levels 25, 50, and 75 on a 100-point scale. Figure 3 shows a typical display and rating page. The whole survey could be completed in approximately 10 minutes, and therefore user tedium should not have had a significant impact on 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 that correlate with the grades as assessed by clinicians. The survey on the Web site permitted respondents to describe the criteria by which they passed judgment; however, even relatively precise statements such as “aver-

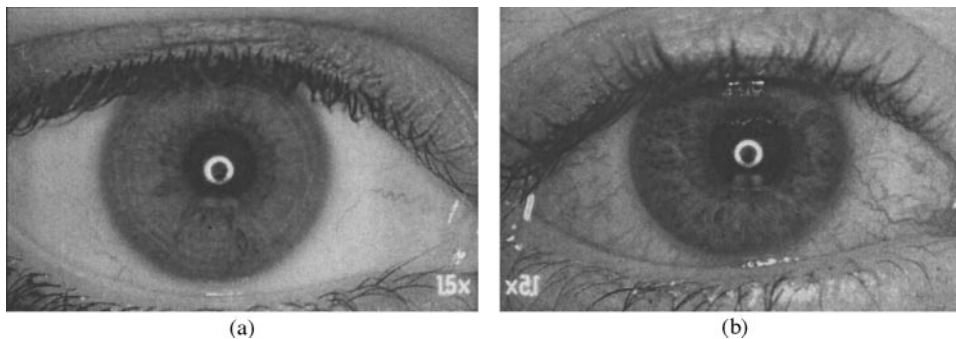
age artery width” or “average redness” are not readily represented as an image-processing algorithm because of the vast number of subjective and subconscious operations undertaken by the human visual system.

Instead, we propose to analyze redness on the basis of two straightforward features, based on a model of the trauma mechanism. Conjunctival hyperemia is characterized 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<sup>11</sup>) to measure the total length of visible arteries. However, the smallest arteries are resolvable neither by the pixels in a charge-coupled device (CCD) camera, nor by the human eye, and a mild onset of hyperemia therefore begins as a diffuse reddening with no discernible edges. In such cases, we propose an integrated measure of redness.

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<sub>R</sub>*, *I<sub>G</sub>*, *I<sub>B</sub>*; for red, green, and blue), is segmented into two nasal and temporal subsets (*S<sub>v</sub>*, *S<sub>n</sub>*, respectively) to allow the two sides to be analyzed separately. The redness feature

$$f_r(S) = \frac{1}{|S|} \sum_{i \in S} \frac{2(S_{Ri}) - (S_{Gi}) - (S_{Bi})}{2[(S_{Ri}) + (S_{Gi}) + (S_{Bi})]} \quad (1)$$



**FIGURE 2.** The best (a) and worst (b) samples of the 30-image set.

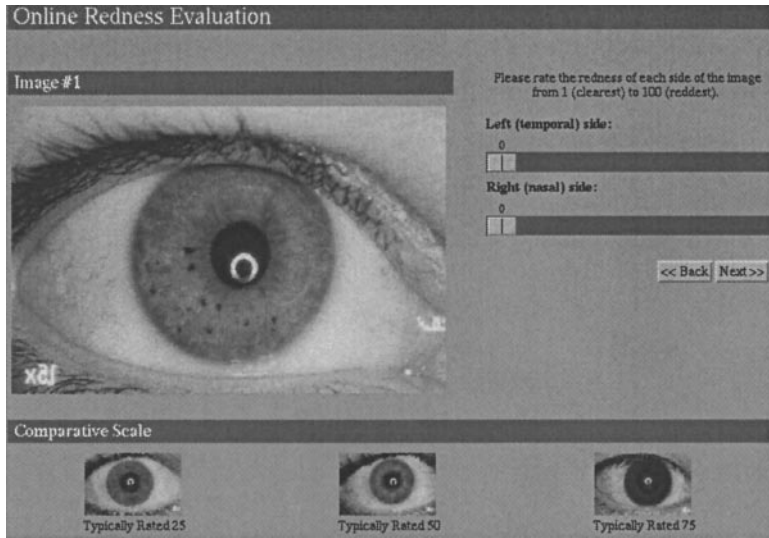


FIGURE 3. A screen shot from the survey web site. The three benchmark images are always visible to the user at the *bottom* of the screen. The two grades (temporal and nasal) are set using the sliders on the *right*.

represents the average integrated redness in the subimage ( $S$ ). Note that black pixels, which have no defined color, have been removed from  $S$ . The denominator term normalizes the feature, so that  $-0.5 < f_r < 1$

$$\begin{aligned} \text{Red image} &\leftrightarrow f_r = 1 \\ \text{White, gray image} &\leftrightarrow f_r = 0 \\ \text{Blue, green image} &\leftrightarrow f_r < 0 \end{aligned} \tag{2}$$

where  $f_r$  is the redness feature. Even the most seriously traumatized 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

$$f_e(S) = \frac{\sum_{i \in S} [\text{Canny}(S)]_i}{|S|} \tag{3}$$

returns the fraction of pixels that 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 color, but is also the number or density of arteries.

In preparing the segmented subsets  $S_n, S_r$ , care had to be taken to eliminate skin pixels with reddish hue that 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 color of the sclera is quite distinct from its surroundings.

## RESULTS

We collected two sets of data: the grades from clinicians through the survey and the computed values of the redness ( $f_r$ ) and edge ( $f_e$ ) features.

### Grading Analysis

Each of the 30 different eye images was graded by 72 clinicians. Although the results show a broad degree of consensus, there was astonishing variability from person to person, shown in Figure 1. The average range in the grades was 55—more than half the entire range of the scale. Even the three calibra-

tion images were 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, because the variability was not observed to increase with the position (early versus late) in the survey.

The histogram of the distribution of assigned grades around the median is shown in Figure 4. The distribution 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 round 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 SD of the grading distribution for each eye varies between approximately 6% and 15%. The trimmed SD, based on keeping only the 50% most consistent grades, is considerably tighter (as implied in Fig. 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  $f_r$ ), and may have varying degrees of consistency (for example, the three or four outliers in the edgeness data in Fig. 6).

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

$$0 \leq \hat{g}(f) \leq 100. \tag{4}$$

The solid lines in the Figures 5 and 6 represent the chosen regressions. Because of the wide range of  $f_r$  (up to 1.0), a linear fit is inappropriate, and a hyperbolic regression was therefore 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

$$g^2(f_r) - 900\hat{g}(f_r) \cdot f_r - 109\hat{g}(f_r) + 90,000f_r - 270 = 0. \tag{5}$$

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

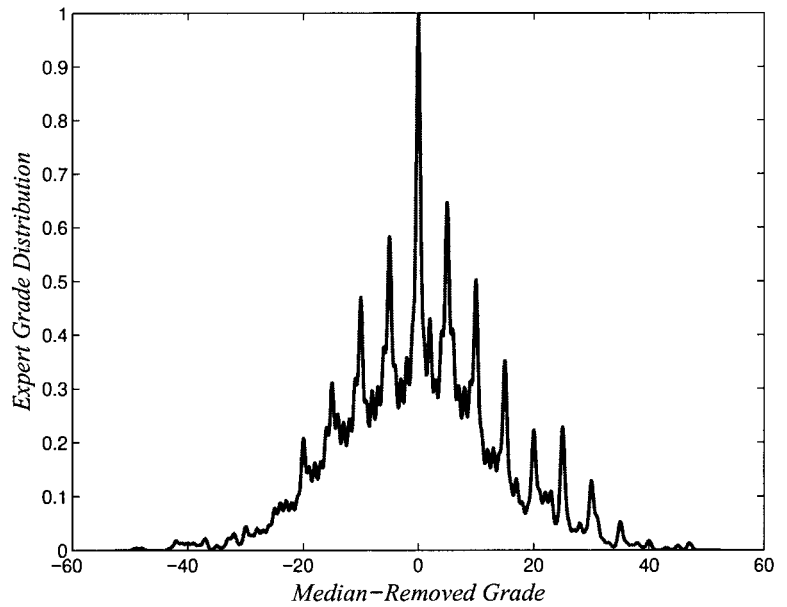


FIGURE 4. The median-removed distribution of grades. The distribution is approximately Gaussian, with an odd periodicity, due to the human bias toward round numbers (multiples of 5 and 10).

A more straightforward linear regression was chosen for  $f_e$ , where the three misfitting data points were eliminated from the coefficient learning process. The resultant expression for the nasal edgeness regression is

$$\hat{g}(f_e) = 5 + 60 \cdot f_e / 0.16. \tag{6}$$

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  $\hat{g}(f_r)$ ,  $\hat{g}(f_e)$  are viewed as approximate “measurements” of the true grade  $g$ , then under certain conditions the optimal linear Bayesian estimate of the grade is

$$\hat{g}(f_r, f_e) = \frac{\hat{g}(f_r)\sigma_e^2 + \hat{g}(f_e)\sigma_r^2}{\sigma_e^2 + \sigma_r^2} \tag{7}$$

and the associated estimation error variance is

$$\text{var}[\hat{g}(f_r, f_e)] = \frac{1}{\sigma_e^{-2} + \sigma_r^{-2}} \tag{8}$$

where  $\sigma_e^2$ ,  $\sigma_r^2$  are the error variances of the single-feature estimators  $\hat{g}(f_e)$ ,  $\hat{g}(f_r)$  respectively. (Ideally,  $\hat{g}(f_r)$ ,  $\hat{g}(f_e)$  should be unbiased estimates of grade  $g$ , and the errors in the two estimates are assumed to be independent.) These error variances cannot be deduced theoretically, but have to be inferred from the data. We computed them as the smoothed local sample variance of the human grades around the regressed curves. The resultant 1-SD curves are shown in Figures 5 and 6. Clearly the Bayesian estimator<sup>7</sup> biases in favor of estimator  $\hat{g}(f_e)$  for eyes having only mild redness, and toward  $\hat{g}(f_r)$  for severe redness.

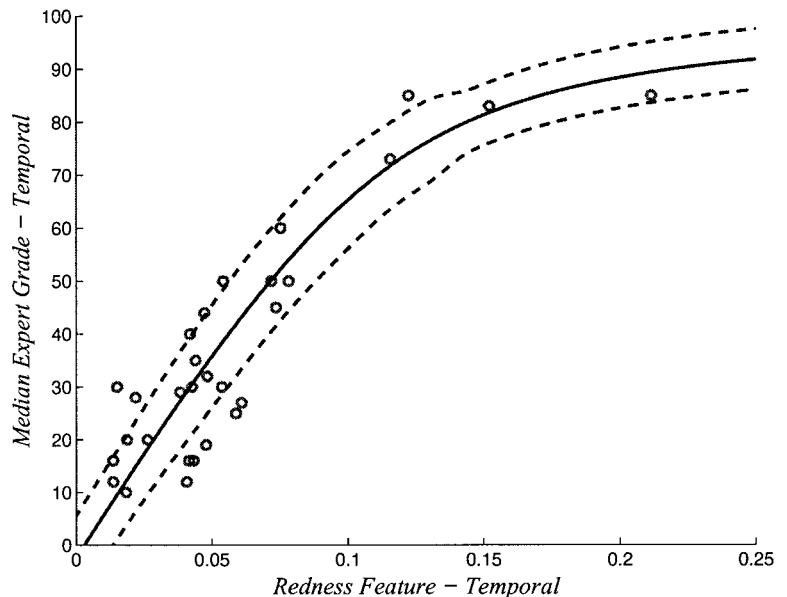


FIGURE 5. Temporal image grades plotted against the integrated redness feature ( $f_r$ ). Solid line: hyperbolic regression (equation 5); dashed lines: the empirical 1-SD envelope around the regression.

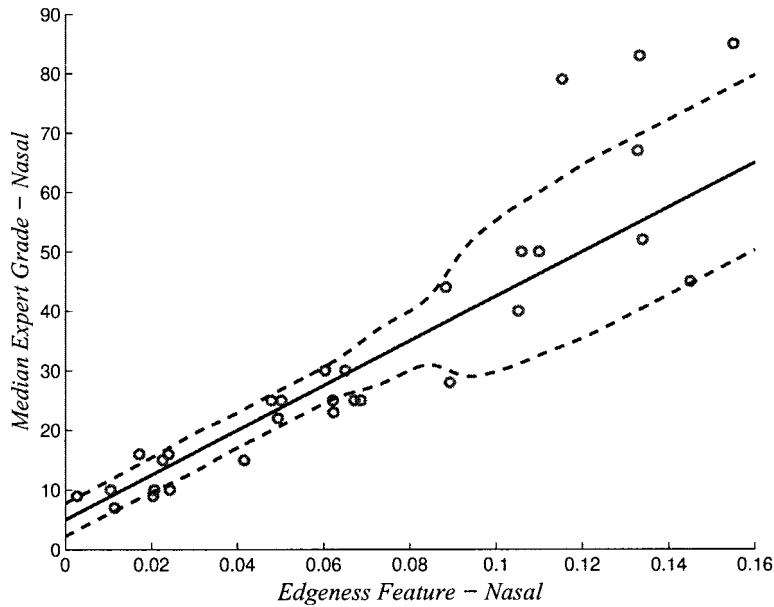


FIGURE 6. Nasal image grades plotted against the edge fraction feature ( $f_e$ ). *Solid line*: linear regression (equation 6), computed omitting the outlying images at large grades. *Dashed lines*: empirical 1-SD envelope around the regression.

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

Figure 7 shows the estimation results, using the Bayesian estimator<sup>7</sup> for both the temporal and nasal data. The estimation results lie very close to the dashed-line 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 SD in length, based on the Bayesian error variance.<sup>8</sup> If the error variances are accurate, they should meaningfully reflect the distribution of the estimates  $\hat{g}$  around the true value  $g$ —that is,

$$\frac{\hat{g}(f_r, f_e) - g}{\sqrt{\text{var}[\hat{g}(f_r, f_e)]}} \tag{9}$$

should be zero-mean, unit-variance Gaussian. Experimentally, the distribution in equation 9 was found to be approximately Gaussian, with a mean of  $-0.04$  and a variance of 1.01, clearly validating the estimated error variances.

Figure 9 compares the error SDs associated with the grading estimates of individuals, the 50% most consistent individuals, and our proposed automated system. Our system represents a great reduction in error over the individuals and except for cases of severe redness, where our regression and learning have a paucity of data, our errors are competitive with the 50% set. Finally, Figure 10 shows the performance of each individual, compared with our proposed system. Of the 72 clinicians who took part in the experiment, only 1 was able to match the

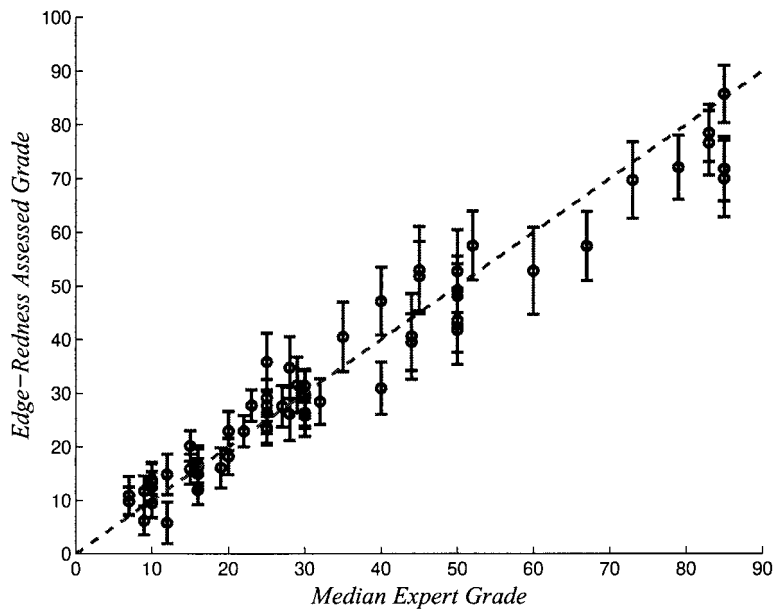


FIGURE 7. Estimator performance for both nasal and temporal data; the correlation coefficient of the fit is 0.976. The fit is best for low grades, for which the most data were available.

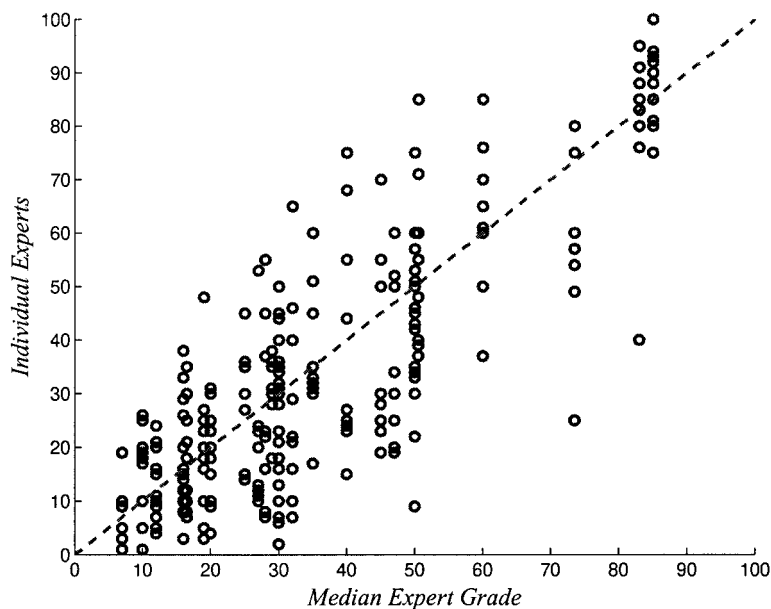


FIGURE 8. Clinical grades, plotted in the same manner as in Figure 7. Clearly, the automated approach yields considerably improved repeatability and error variances.

consistency (measured as the correlation coefficient) of our proposed method. Clearly, our errors are competitive with or better than even the most consistent graders.

**DISCUSSION**

The results of this study reinforce previous results. Automated measures provide information that is linearly associated with subjective grades of redness. Our results are similar to those of Willingham et al.<sup>6</sup> and Papas,<sup>9</sup> in that we each found strong associations between the subjective grades and the measurements, as opposed to the weaker associations of Guillon and Shah.<sup>5</sup> Our methodology is more similar to that of Willingham et al. and Papas, inasmuch as we used images that were graded, whereas Guillon's and Shah's subjective data were collected in vivo with a slit lamp biomicroscope. Our methods differ from those of 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 features that is remarkably concordant with the grades assigned by clinicians.

A primary objective of this study was to minimize the required operator intervention. In some previous studies combinations of custom software and hardware have been used, making the analysis inaccessible or expensive to develop, operate, and maintain. For our research, commonly used desktop computers (Pentium processor; Intel, Mountain View, CA; running Windows; Microsoft, Redmond, WA) performed the data acquisition, and numeric processing was performed in an environment based on a widely available program (Matlab; MathWorks, Natick, MA). 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.

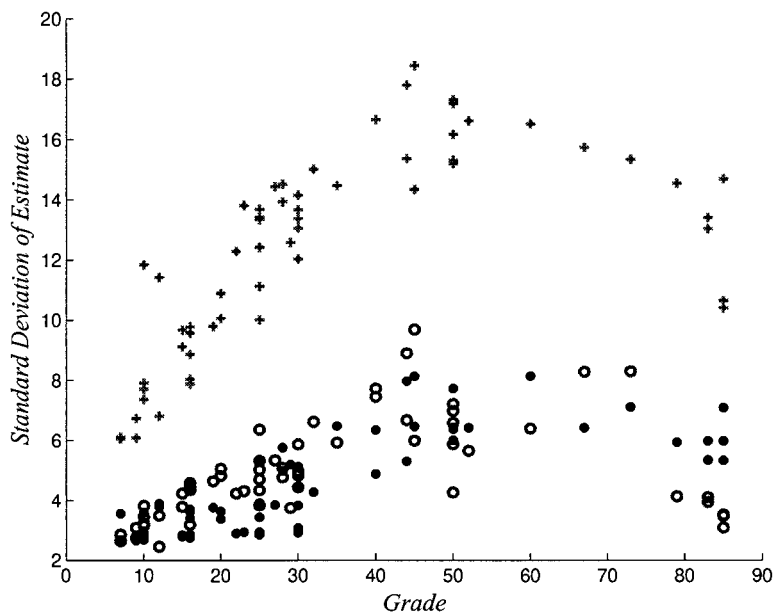
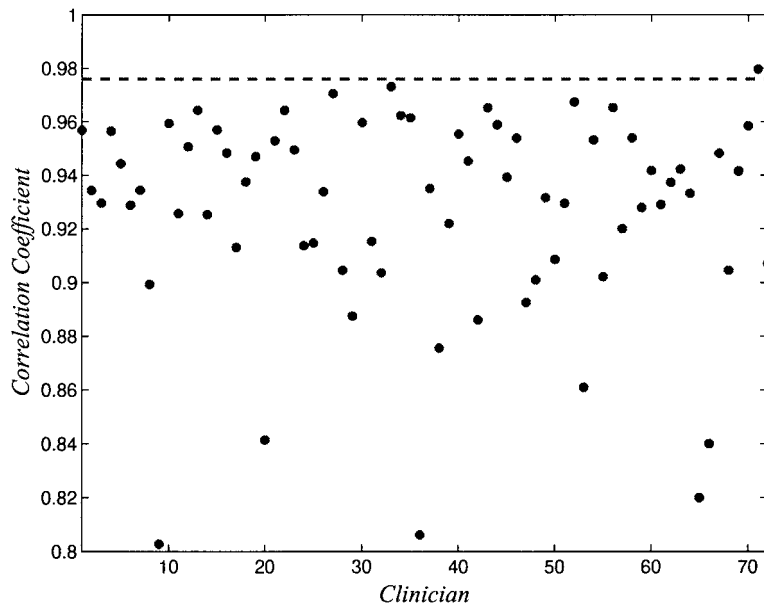


FIGURE 9. SDs of the estimation errors: (+) clinician, (○) the 50% most consistent clinicians, and (●) the proposed automated system. Except for grades at the severe end of the scale, where the data were sparse, the errors of the proposed system are as good as or better than the 50% consensus of a group 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. *Dashed line:* correlation of 0.976 attained by the automated system.

These processing steps can therefore 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 observers and that our sampling method was far from randomized. These experimental attributes are no different, however, from any of the previous reports comparing automated methods to subjective methods of grading. Our sampling method provided additional diversity, in that the clinicians 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 graders were true experts who used 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.<sup>12</sup> This was not accomplished by accident, because the graders would have had 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,<sup>13,14</sup> but the exact impact on the accuracy and repeatability of redness grading would have to be determined empirically.

Another result relating to grading was the large range of grading associated with the reference images that were part of the data set that clinicians graded. Although the median grades were very similar to the reference grades (Fig. 1), there was a surprisingly large range of grades associated with each reference. This suggests that the clinicians either could not psychophysically match grades to the references, an unlikely conclusion, or that they chose to ignore the values assigned to the reference. Clinicians have been shown to resist using tools that are assistive,<sup>15</sup> 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 were 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 are illustrated in Figure 8, using resampling techniques.<sup>16</sup> 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 SD of the computed approach for each slide with the actual SD 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 telemedicine applications? Image compression reduces storage or transmission needs but may also be associated with a loss of information. All the images in our analysis (and web survey) were stored in a lossless (tagged information file format [TIFF] file) 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 compression may be developed that minimizes the loss of critical information at the expense of unimportant image content—for example, by examining the changes in the image features  $f_r, f_e$  as a function of the compression type. This is similar to previous suggestions<sup>17</sup> using perception to constrain image coding, except that the data are clinically salient, rather than only perceptually salient.<sup>8</sup>

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 compared very well with those derived using clinical grading methods. In addition, the method we propose has much less variability than that which exists between clinical graders. In the past 10 years, numerous methods have been proposed to assign redness scores by computational methods. The question then might be, how does the procedure developed here advance these methods? For example, Villumsen et al.<sup>8</sup> 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 and software combina-

tions, the algorithms we used are available to anyone using a computer running almost every operating system that may be encountered. 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.

### Acknowledgments

The authors thank all the observers who visited the web site and completed the grading task; the following former students for their dedicated efforts in preparing the survey web site and collecting the grading data: Janine Cullen, Cynthia Handler, Vikas Nagaraj, Shannon Nichols, Shane Pounder, and Kimberly Whitear; and the anonymous reviewers who provided helpful suggestions for improvement of the manuscript.

### References

1. Steven SS. On the theory of scales of measurement. *Science*. 1946; 103:677-680.
2. McMonnies C, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. *Am J Optom Vis Sci*. 1987;64: 246-250.
3. Chong T, Simpson T, Fonn D. The repeatability of discrete and continuous anterior segment grading scales. *Optom Vis Sci*. 2000; 77:244-251.
4. Cullen J, Pounder S, Whitear K, Fieguth P. *Analysis of Corneal Images for Assessing Contact Lens Trauma*. International Conference on Image Processing. Vancouver, British Columbia, Canada, 2000.
5. Guillon M, Shah D. Objective measurement of contact lens induced conjunctival redness. *Optom Vis Sci*. 1996;73:595-605.
6. Willingham FF, Cohen KL, Coggins JM, Tripoli NK, Ogle JW, Goldstein GM. Automatic quantitative measurement of ocular hyperemia. *Curr Eye Res*. 1995;14:1101-1108.
7. Owen CG, Fitzke FW, Woodward EG. A new computer assisted objective method for quantifying vascular change of the bulbar conjunctiva. *Ophthalmic Physiol Opt*. 1996;16:430-437.
8. Villumsen J, Ringquist J, Alm A. Image analysis of conjunctival hyperemia: a personal computer based system. *Acta Ophthalmol*. 1991;69:536-539.
9. Papas E. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci*. 2000;41:687-691.
10. Terry RL, Schnider CM, Holden BA, et al. CCLRU standards for success of daily and extended wear contact lenses. *Optom Vis Sci*. 1993;70:234-243.
11. Castleman K. *Digital Image Processing*. Englewood Cliffs, NJ: Prentice Hall; 1996.
12. Efron N, Morgan P, Katsara S. Validation of grading scales for contact lens complications. *Ophthalmic Physiol Opt*. 2001;21: 17-29.
13. Bailey IL, Bullimore MA, Raasch TW, et al. Clinical grading and the effects of scaling. *Invest Ophthalmol Vis Sci*. 1991;32:422-432.
14. Schulze M. The production of an enhanced grading scale for determination of ocular hyperaemia. *University of Waterloo Technical Report*. Waterloo, Ontario Canada: University of Waterloo; 2000.
15. Dawes R, Faust D, Meehl P. Clinical versus actuarial judgment. *Science*. 1989;243:1668-1674.
16. Davison AC, Hinkley DV. *Bootstrap Methods and Their Application*. Cambridge, UK: Cambridge University Press; 1997.
17. Chou C-H, Li Y-C. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. *IEEE Trans Circuits Syst Video Technol*. 1995;5:467-476.