tvst

https://doi.org/10.1167/tvst.8.6.31

Article

Validation of Computerized Quantification of Ocular Redness

Ekaterina Sirazitdinova[1], Marlies Gijs[2], Christian J. F. Bertens[2], Tos T. J. M. Berendschot[2], Rudy M. M. A. Nuijts[2,3], and Thomas M. Deserno[4]

[1] Uniklinik RWTH Aachen, Department of Medical Informatics, Aachen, Germany
[2] University Eye Clinic Maastricht, Maastricht University Medical Center+ (MUMC+), Maastricht, the Netherlands
[3] Department of Ophthalmology, Zuyderland Medical Center, Heerlen, the Netherlands
[4] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany

Correspondence: Thomas M. Deserno, PLRI, Mühlenpfordtstr. 23, 38106 Braunschweig, Germany. e-mail: thomas.deserno@plri.de

Received: 19 November 2018
Accepted: 20 October 2019
Published: 12 December 2019

Keywords: ocular redness; conjunctival provocation test; allergy; image processing; clinical trials; clinical grading; sclera segmentation

Citation: Sirazitdinova E, Gijs M, Bertens CJF, Berendschot TTJM, Nuijts RMMA, Deserno TM. Validation of computerized quantification of ocular redness. Trans Vis Sci Tech. 2019;8(6):31, https://doi.org/10.1167/tvst.8.6.31

Purpose: To show feasibility of computerized techniques for ocular redness quantification in clinical studies, and to propose an automatic, objective method.

Methods: Software for quantification of redness of the bulbar conjunctiva was developed. It provides an interface for manual and automatic sclera segmentation along with automated alignment of region of interest to enable estimation of changes in redness. The software also includes the redness scoring methods: (1) contrast-limited adaptive histogram equalization (CLAHE) in red-green-blue (RGB) color model, (2) product of saturation and hue in hue-saturation-value (HSV), and (3) average of angular sections in HSV. Our validation pipeline compares the scoring outcomes from the perspectives of segmentation reliability, segmentation precision, segmentation automation, and the choice of redness scoring methods.

Results: Ninety-two photographs of eyes before and after provoked redness were evaluated. Redness in manually segmented images was significantly different within human observers (interobserver, $P = 0.04$) and two scoring sessions (intraobserver, $P < 0.001$). Automated segmentation showed the smallest variability, and can therefore be seen as a robust segmentation method. The RGB-based scoring method was less sensitive in redness assessment.

Conclusions: Computation of ocular redness depends heavily on sclera segmentation. Manual segmentation appears to be subjective, resulting in systematic errors in intraobserver and interobserver settings. At the same time, automatic segmentation seems to be consistent. The scoring methods relying on HSV color space appeared to be more consistent.

Translational Relevance: Computerized quantification of ocular redness holds great promise to objectify ocular redness in the standard clinical care and, in particular, in clinical trials.

Introduction

A wide range of ocular conditions are characterized by bulbar redness including dry eye disease, (allergic) conjunctivitis, blepharitis, corneal abrasion, foreign body, subconjunctival hemorrhage, keratitis, iritis, glaucoma, chemical burn, and scleritis.[1] In addition, ocular redness is often observed in contact lens wearers.[2] Ocular redness is a sign of ocular inflammation and is generally associated with pain or discomfort and often accompanied with vision problems.

Ocular redness is an important diagnostic feature to detect diseases and to monitor disease progression and treatment. In clinical practice, the most common way to grade eye redness relies on the usage of special reference scales. The most known grading scales are the McMonnies/Chapman-Davies scale,[2] Efron scale,[3] the Institute for Eye Research scale (also known as CCLRU),[4] and the validated bulbar redness

scale.[5] Using such techniques, a clinician grades the patient's condition using photographic[2,4,5] or artist-rendered[3] reference images. This method is very simple, and a trained clinician would need approximately 10 seconds in order to accomplish grading. However, these methods also have several major drawbacks. First, the grading is highly subjective because it depends on the knowledge and experience of the clinician. Secondly, due to the limited set of grading states, it cannot provide continuous linear quantitative evaluation, which makes these methods not very sensitive to small changes in ocular redness in early stages of disease. However, this sensitivity is of high importance for early diagnosis and in clinical trials,[6] which evaluate the safety of new ophthalmic drugs, drug formulations, or drug delivery devices.[7] Furthermore, because of the lack of photographic documentation, grading by this method is not reproducible and does not allow for a second observer. Hence, despite a relatively high number of existing approaches, none of them is regarded as a gold standard.

In the present study, we investigated the reliability of computerized techniques for ocular redness quantification. In particular, we are interested in establishing the reliability of the redness score depending on region of interest (ROI) segmentation and a chosen scoring method. Furthermore, we propose a processing pipeline designed to avoid subjectivity by replacing all human interactions with automated algorithms.

Materials and Methods

In order to extract data from ocular photographs, we developed a software tool featuring a graphical user interface (GUI) for sclera selection and segmentation. After image acquisition, we implemented a machine learning method for automatic sclera segmentation, which is independent of image size, eye pose, and illumination. Based on the concept of Sárándi et al.,[6] a method was developed for the selection of the ROI. ROI registration and intersection was performed in corresponding images using feature matching,[8] assuring that exactly the same part of the eye is considered for the computation of redness scores over the time. For redness scores, we implemented and compared the approaches of Park et al.,[9] Amparo et al.,[10] and Sárándi et al.[6] Figure 1 illustrates our processing pipeline.

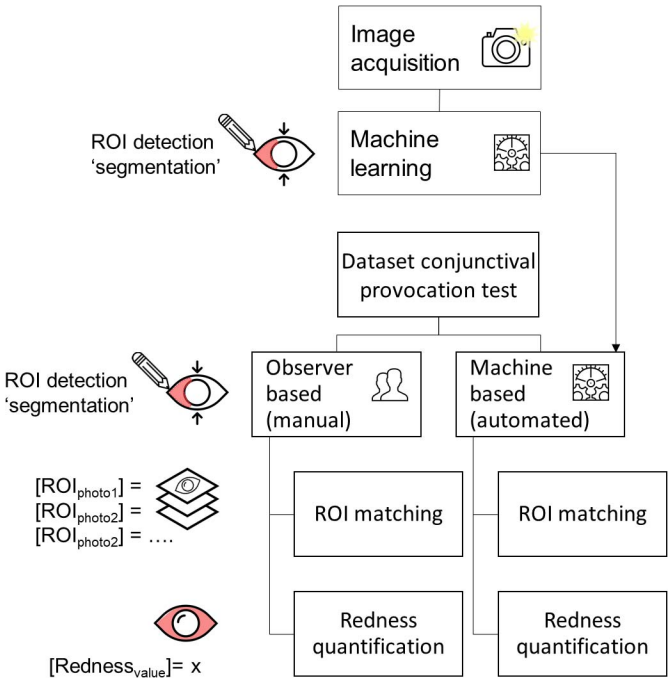


Figure 1. Organizational chart of the experiment.

Image Acquisition

For software development (training of a machine learning classifier) and preliminary testing, a total of 97 photographs of 18 volunteers were taken at the University Eye Clinic Maastricht (Maastricht, the Netherlands). The protocol was approved by the local ethics committee and the national authorities. The study procedures were performed in accordance with the tenets of the Declaration of Helsinki. All participants signed written informed consent before inclusion. Three photographs were taken per eye at 6.3× times magnification using a calibrated Haag-Streit BX900 slit-lamp bio-microscope (Haag Streit AG, Bern, Switzerland) in combination with a computer-operated digital camera (Nikon D7100; Nikon, Tokyo, Japan). The volunteers were asked to look left, right, and up. Images were exported as JPG files (2992 × 2000 pixels, 150 dpi). Background illumination was used on full intensity (100% open), and grey-filter settings were set to 100% open. Slit beam illumination was used with a diffusion filter, a width of 15 and 8 mm height of the beam at a 45° oblique angle.

For evaluation, the data set from the conjunctival provocation test[6] was used. The data set contains 92 images of 23 patients. The images were taken in pairs: before (called "reference image") and after (called "response image") the application of an inducing

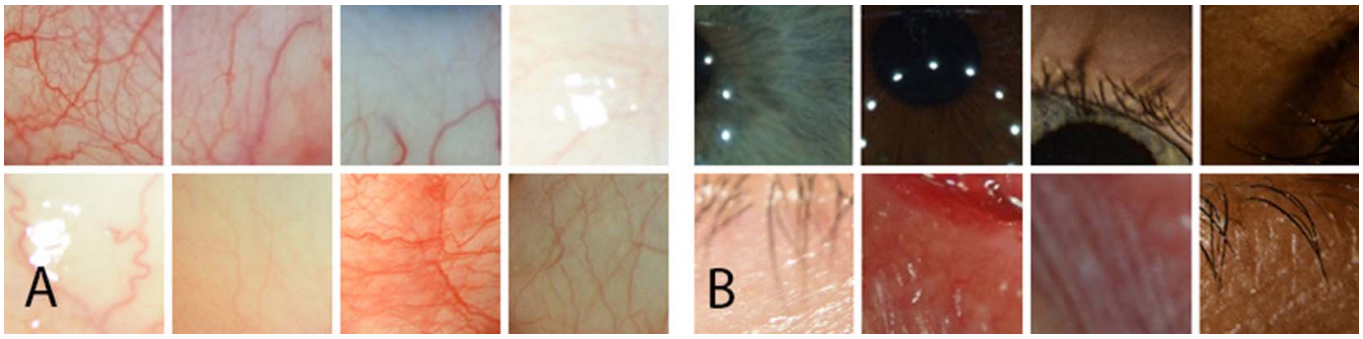


Figure 2. Training patches. (A) Example sclera patches. (B) Example nonsclera (*background*) patches.

redness allergen. For each patient, the procedure was performed twice in separate visits (visit 1 and visit 2). The data set used was recorded in the controlled environment with the same equipment (see Supplementary Fig. S1).

Automatic ROI Detection

For ROI segmentation, nonparametric models (i.e., random decision forest) were used.[11] For training, we used the open-source machine learning software Weka[12] and the Trainable Weka Segmentation (TWS) toolkit.[13] It utilized a fast (i.e., multithreaded) version of Breiman's random forest algorithm.[14] We initialize with 512 "trees" and eight random features per node. These parameters were derived empirically. Images of eight different subjects were used for training: the subjects feature different eye color and skin tone, and level of redness and prominence of vascular structure vary within selected samples. Therefore, two classes of regions were selected manually: sclera and background (Fig. 2). Approximate training time was 10 seconds per image on the used hardware (Intel Core i7-2620m processor, 8 GB RAM).

Classification is integrated in our custom software written in Java. Using a trained model, grayscale probability maps are created for new images where higher intensities correspond to the regions that most likely belong to the sclera (Fig. 3A). Simple postprocessing involving binary threshold and morphological operators is applied to the probabilistic maps such that the largest area with the highest probability score is identified as the ROI (Fig. 3B). The outer contour of the detected ROI is then processed with Bresenham's line algorithm,[15] which smoothens the contour and provides adjustment points, which can be used in the GUI in order to correct the detected ROI manually if necessary (Fig. 3C).

Manual ROI Detection

Five human observers performed manual segmentation using the GUI interface running on the same machine. Four of the human observers performed the segmentation of each image twice. In each manually segmented image pair consisting of the images of the same eye before application of the allergen and after, redness scores were estimated both before and after applying ROI matching.

ROI Matching

If we want to achieve the most precise comparison between different stages of redness in the same eye, the same parts of the sclera on the photographs need to be measured. It is incorrect to compare redness in two ROIs just after ROI detection because of possible differences in eyelid openness, differences in gaze direction, and also different scales and image resolutions associated with nonstandardized acquisition settings. Therefore, we implemented the registration of two or more sequential ROIs to find a common ROI that shall be used for redness computation. The method is based on detection of landmarks, or points of interest, which are robust to rotation, translation, and scale. Scale invariant feature transform (SIFT) points of interest are detected in all ROIs, and point correspondences are estimated by feature similarity.[8] Random sample consensus (RANSAC) is used for robustness refinement.[15] Using these correspondences (Fig. 4A), transformation between the reference and the matched ROIs can be derived and applied to matched ROI. The transformed ROI is laid over the reference ROI, and the intersection of both is used as the common ROI for redness estimation (Fig. 4B). This is also beneficial for removal of false positives in ROIs.

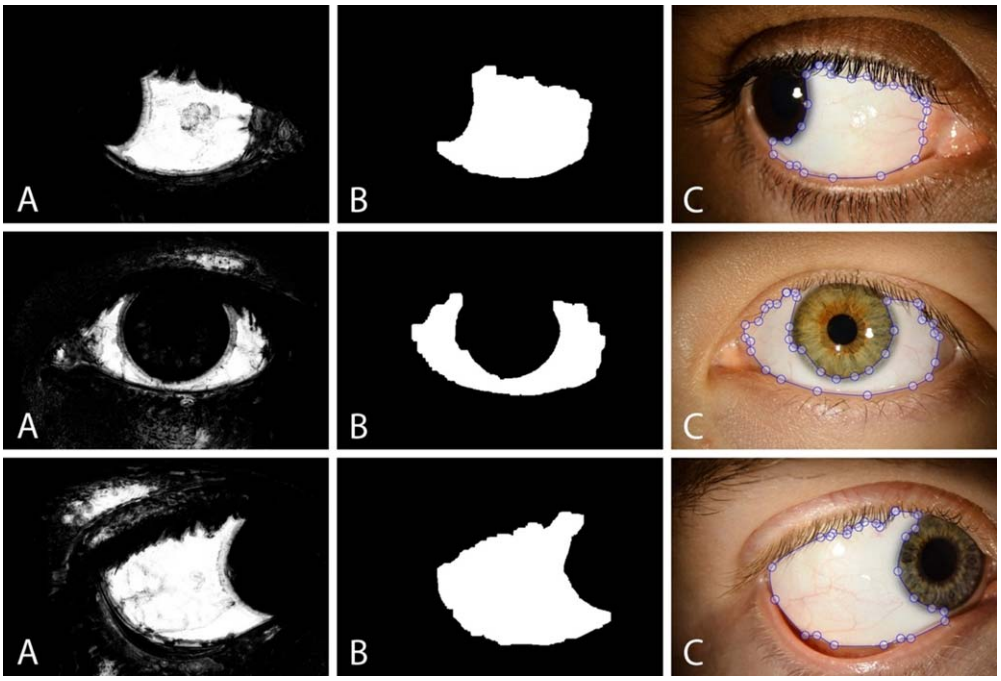


Figure 3. Segmentation steps applied to three different subjects: (A) Generated probability map of sclera segmentation: higher intensities correspond to the areas, which most likely belong to the sclera region. (B) ROI derived out of the probability map using simple thresholding and refinement with morphological operations of erosion and dilation. (C) ROI with adjustment points laid over the original image.

Redness Quantification

For redness scores, we implemented the approaches of three different studies: Park et al.,[9] Amparo et al.,[10] and Sárándi et al.[6] (Fig. 5). Park et al.[9] have used the contrast-limited adaptive histogram equalization (CLAHE) for blood vessels enhancement.[16] The vessels are segmented using thresholding, and the redness score is calculated as a ratio of number of pixels corresponding to the blood vessels to the total number of pixels in the ROI. Amparo et al.[10] use hue-saturation-value (HSV) color space for redness estimation and use the product of saturation and hue mapped to $[0, 1]$ interval as the redness score. Sárándi et al.[6] also rely on HSV color space and compute the redness score as an average of maximal values $\max\{0, S, \cos(2\pi H)\}$ computed for each pixel in the ROI, where S and H are saturation and hue components of the pixel, respectively.

Clinical Cases

To test the final version of the program, three clinical cases of ocular redness were assessed in the University Eye Clinic Maastricht. Patients signed written informed consent before photos were taken. From both eyes three photos were taken using slit-lamp settings as described in Image Acquisition. Patients were asked to glare up, left, and right.

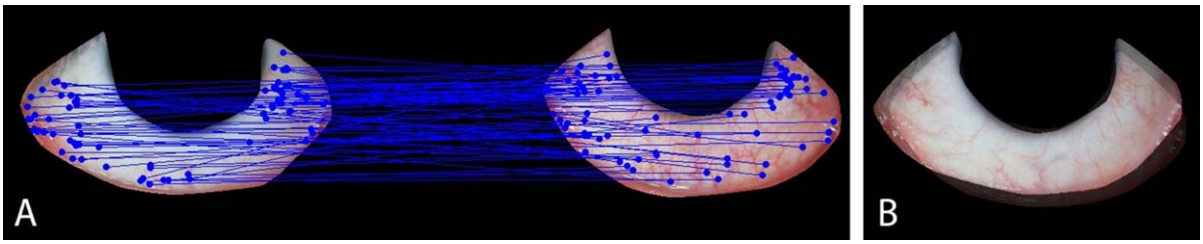


Figure 4. (A) Corresponding points of interest are connected with *straight lines*. Photo on the *left* was taken before the allergen was applied, and on the *right* — after. (B) Overlay of registered ROIs: only the overlapping area is considered as ROI for redness computation.

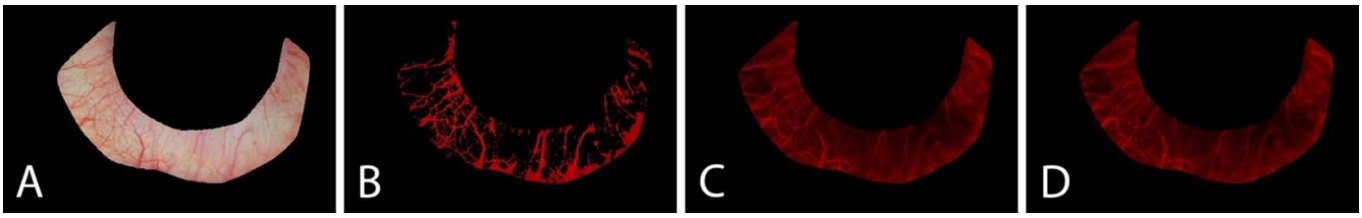


Figure 5. (A) ROI selected in the original image. Pixels classified as red using the methods of (B) Park et al.,[9] (C) Amparo et al.,[10] (D) Sárándi et al.[6]

Statistical Analysis

The described system was utilized to determine redness scores computed using three different methods (Park et al.,[9] Amparo et al.,[10] and Sárándi et al.[6]). First, *segmentation reliability*, defined as the ability of the observer to produce similar results time after time, also known as intraobserver difference, was evaluated using a test-retest fashion (Bland and Altman plot). To estimate the significance level of difference in redness scores within test and retest segmentations, mean reference and response redness values of both visits were compared using a paired *t*-test and a general linear model repeated measures test. To exclude the effect of other features, no ROI matching was performed, and the score was computed.[6] Second, *segmentation precision* was defined as interobserver difference. For that, the mean redness values of the reference recordings in the first visit, using test only (the first segmentation by five human observers), were compared using a general linear model repeated measures test. Again, no ROI matching was performed, and the score was computed.

To estimate the robustness of a computer-based method, we evaluated the effect of *segmentation automation* by comparing the differences between visit 1 and visit 2 of the reference images between values computed with and without ROI matching using analysis of variance (ANOVA). In addition, to prove the assumption that if we include ROI registration, the absolute redness values indicating changes in redness are supposed to be more robust, we computed the scores with and without applying the proposed technique.

We implemented three *redness scoring methods* and, based on the assumption that a large difference in redness between reference and response shall indicate higher sensitivity, we compared redness differences between reference and response values estimated by all three methods (Park et al.,[9] Amparo et al.,[10] and Sárándi et al.[6]) using automatic segmentations provided by our machine learning method (without ROI matching).

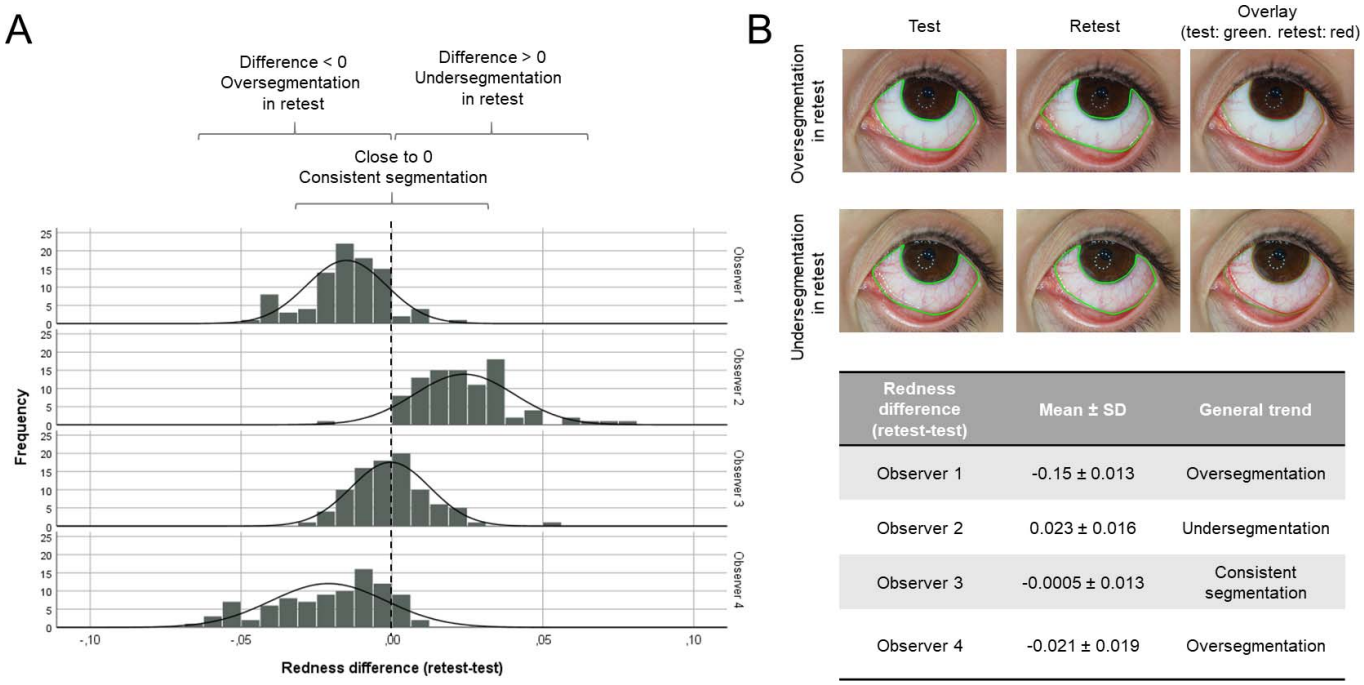
In order to illustrate the clinical applicability by case, we selected three trial subjects from the conjunctival provocation test panel. Based on the subjective assessment on visual differences between the reference image and response image, these subjects were labelled as strong, mild, or no responders to the provocation test.

All data are analyzed using SPSS (version 25 IBM, Armonk, NY), and data are shown as mean ± standard deviation (SD).

Results

Segmentation Reproducibility

There was a significant difference ($P < 0.001$) between the test and retest for three out of four human observers (Fig. 6A), meaning that there was a systematic error for the three observers. Further, these systematic errors differed between the observers ($P < 0.001$). Frequency distributions of differences in redness scores between test and retest observations (Fig. 6B) indicate that segmentation by observers 1 and 4 systematically results in larger redness values during the retest ("oversegmentation"), while segmentation by observer 2 systematically provides smaller redness values ("undersegmentation"). Observer 3 is consistent in his manual segmentation. Additionally, observers 2 and 4 display a broad variability in redness values in contrast to observers 1 and 3. These trends are illustrated by two case examples of oversegmentation and undersegmentation and by their mean values of redness difference. Figure 7 shows the differences between test and retest versus the mean grading estimate. There is no general relation between the differences and the means, indicating that segmentation reliability is unaffected by the redness score itself. Again, observer 3 shows the best segmentation reliability as a tighter cluster of redness differences around zero can be recognized,



Redness difference (retest-test)	Mean ± SD	General trend
Observer 1	-0.15 ± 0.013	Oversegmentation
Observer 2	0.023 ± 0.016	Undersegmentation
Observer 3	-0.0005 ± 0.013	Consistent segmentation
Observer 4	-0.021 ± 0.019	Oversegmentation

Figure 6. (A) Frequency distributions of redness differences between test and retest observations for four human observers. (B) Example of test and retest with an overlay from an oversegmentation and an undersegmentation. The table shows an overview of the general trend from the observers.

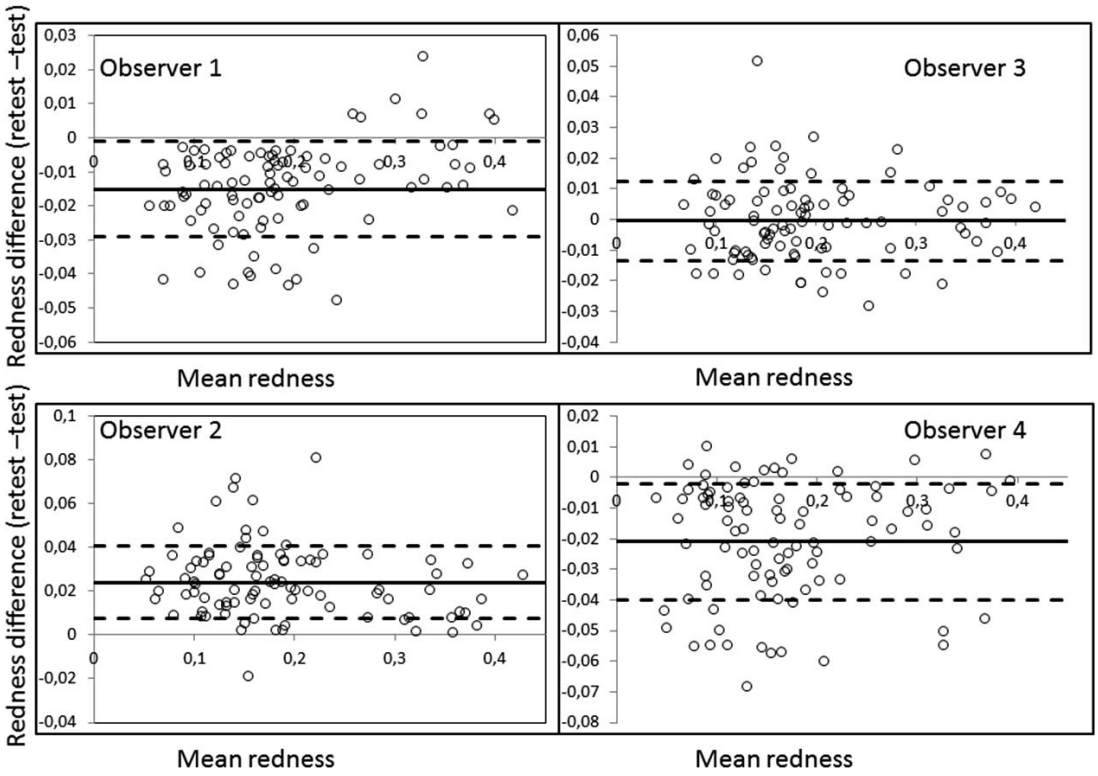


Figure 7. Redness difference versus mean redness of test and retest redness values for four human observers. The *thick solid line* represents the mean value of test-retest discrepancies, and the *dotted lines* represent the mean ± SD.

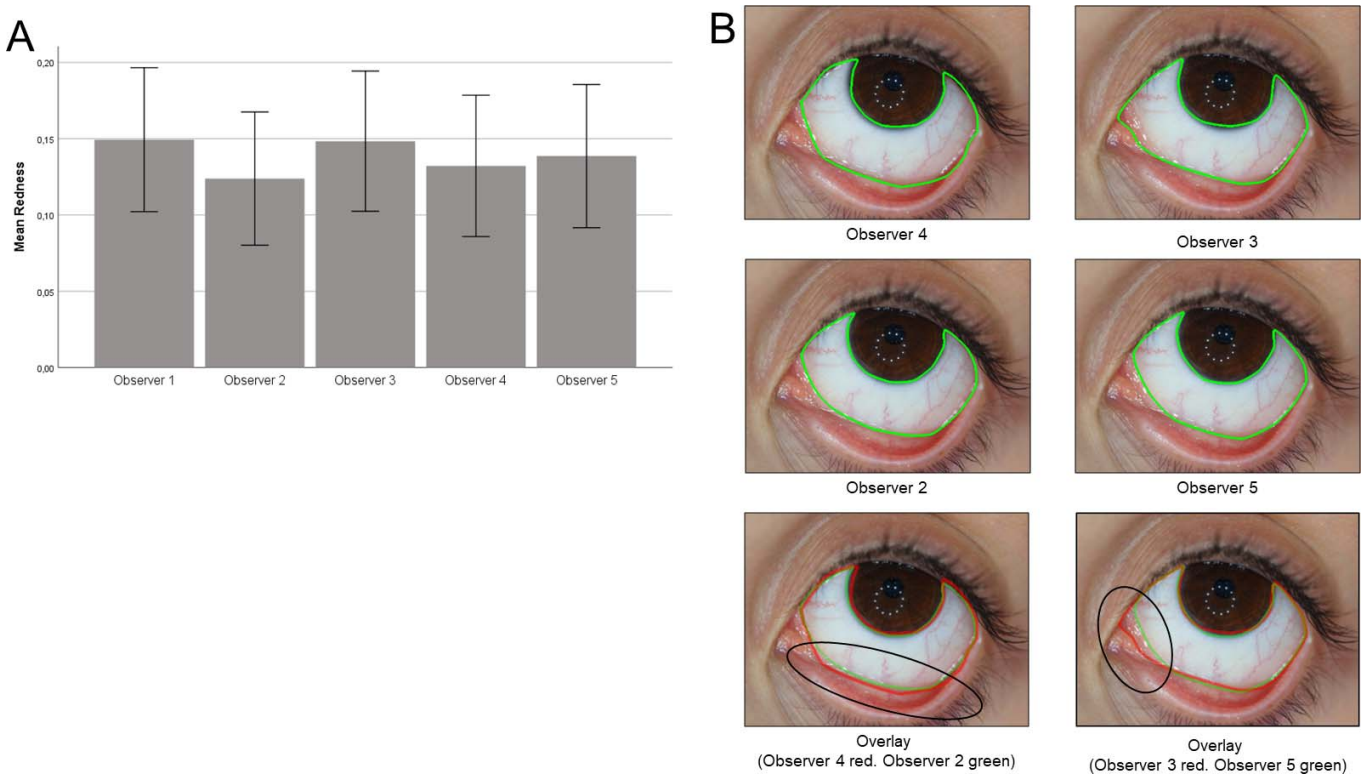


Figure 8. (A) Mean redness values (± SD) of the reference recordings in the first visit, using test only, without ROI matching, computed using the method of Sárándi et al.[6] for five observers. (B) Differences between observers related to the conjunctival border (*left column*) and the semilunar conjunctival fold (*right column*).

while more values falling far from the mean are seen for observers 1, 2, and 4.

Segmentation Accuracy

The interobserver difference, that is, the difference between multiple human observers for the reference images, was significantly different between the five observers ($P = 0.040$) (Fig. 8A) meaning that manual segmentation is easily affected by subjective factors (Fig. 8B).

Segmentation Automation

The overall mean redness difference of the human observers showed an increase by implementing ROI matching, however insignificant (Figs. 9A, 10). This is illustrated by two case examples segmented by observer 4 that shows an increase in redness difference after implementation of ROI matching (Fig. 9B). With the machine learning approach, ROI matching improved the results as the mean redness difference became smaller, though insignificant as well.

Redness Scoring Method

Figure 11 shows that the redness values calculated by the method of Park et al.[9] largely overlap and, thus, is insufficiently able to detect differences in redness. In contrast, little overlap can be observed at the methods by Amparo et al.[10] and Sárándi et al.[6] The sensitivities of these two methods are similar. Three case examples illustrate that the method of Park et al.[9] is insensitive to detect differences in redness for the strong and mild responder, while the sensitivities of Amparo et al.[10] and Sárándi et al.[6] are comparable (Fig. 12).

Clinical Application by Case

Our automated tool generated nominal values of redness difference between the reference (before) and response (after) images (Fig. 13). Although the subjective assessment in these simplistic examples is straightforward, one can appreciate the sensitivity of our automated tool, with up to nine-fold differences in redness difference between two cases of the same participant.

When no follow-up visit is available, redness can

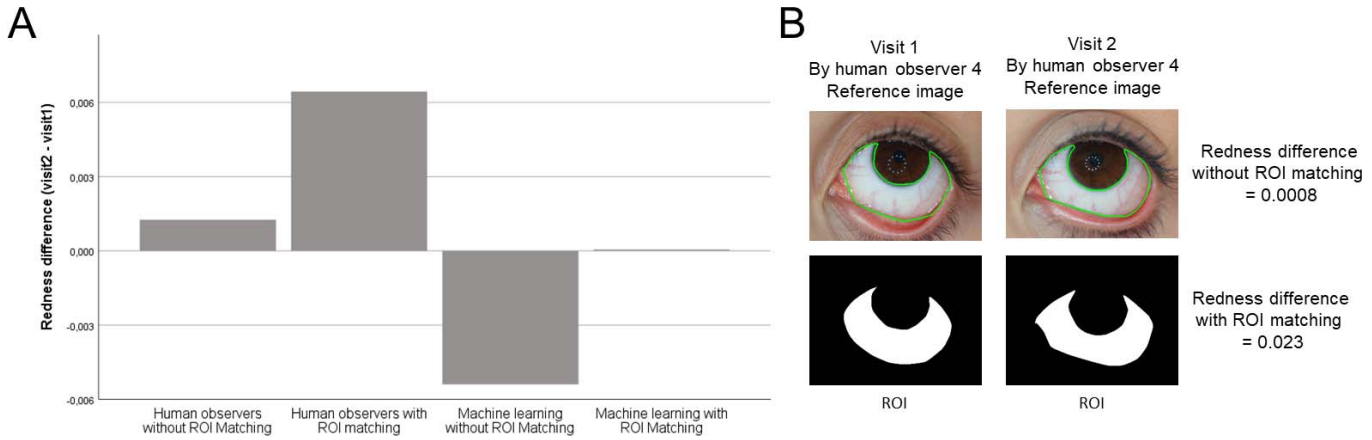


Figure 9. (A) Mean redness differences between visit 1 and visit 2 of the reference images for all human observers and the machine-learning approach, both with and without ROI matching. (B) Example of redness difference with or without ROI matching.

be scored using the contralateral eye as shown in Figure 14A. Three clinical cases are tested using the methods by Park et al.,[9] Amparo et al.,[10] and Sárándi et al.[6] (Fig. 14B). In all methods, the affected eye provides a higher redness value compared with the contralateral eye. The values generated by the methods of Amparo et al.[10] and Sárándi et al.[6] are almost two times higher in intensity compared with the values generated by the method of Park et al.[9]

Discussion

Ocular redness is an observable clinical response of the ocular surface in pathological conditions. To some extent, the degree of redness may reflect the severity of the disease. In this context, quantification of ocular redness can be of use in both clinical and research settings. Examples of conditions that are often associated with ocular redness are dry eyes

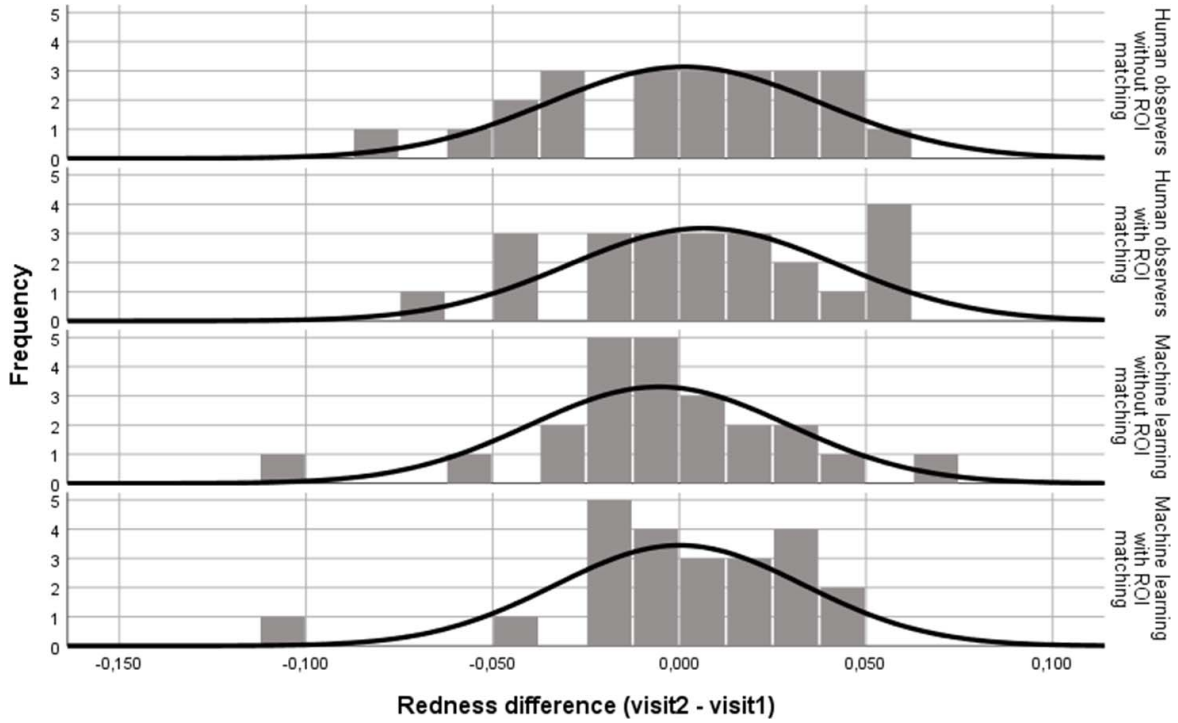


Figure 10. Frequency distribution of the redness differences between visit 1 and visit 2 for all human observers and the machine-learning approach, both without and with ROI matching.

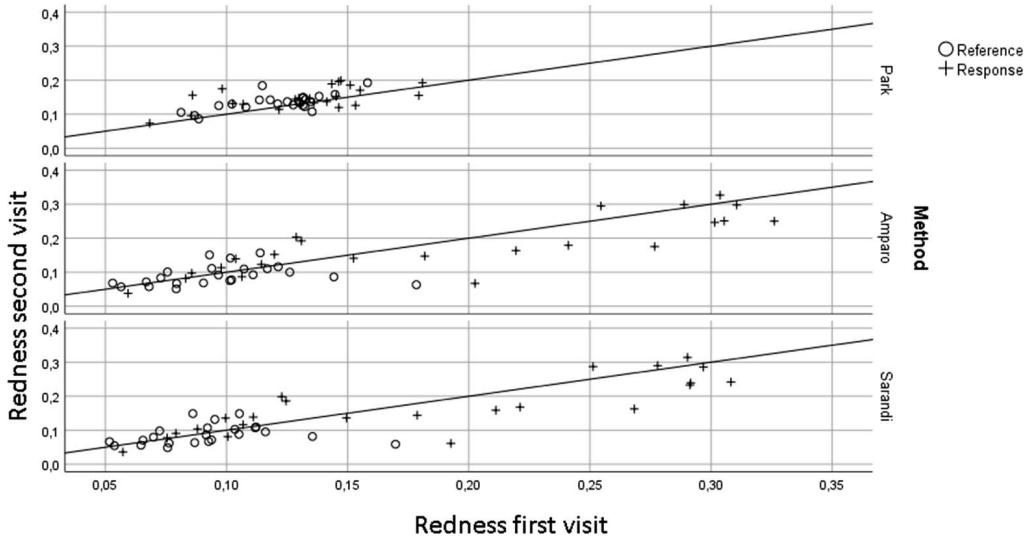
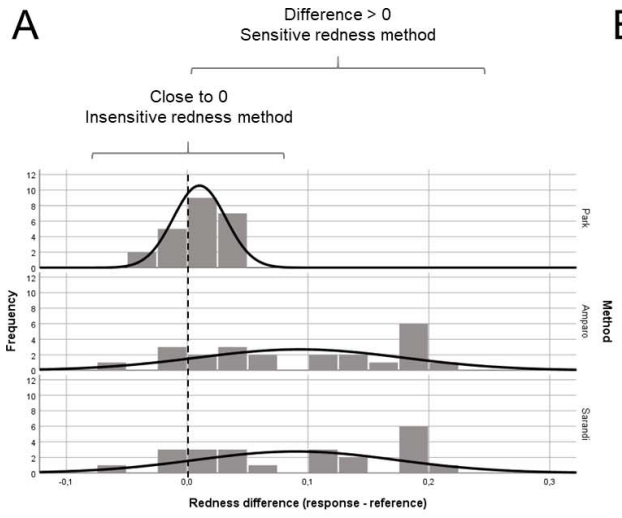


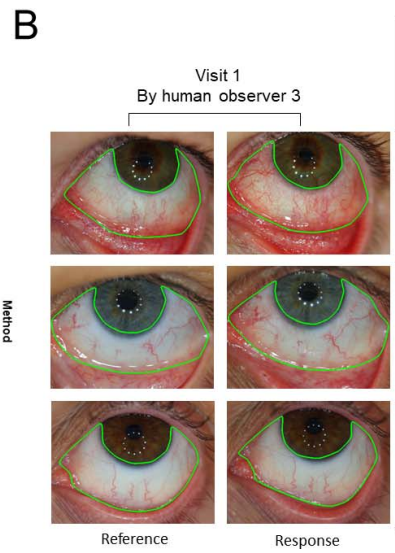
Figure 11. Comparison of redness scores for the machine-learning approach of three different redness scoring methods without ROI matching. The *solid line* shows the equality.

disease, contact lens complications, and allergic conjunctivitis. In clinical practice, sensitive quantification of ocular redness would allow to stage the (subclinical) disease, to monitor progression of the disease and to control and regulate treatment efficacy.

Another application for computerized quantification of ocular redness would be in a setting of multicenter clinical trial to investigate the safety of new topical drugs or devices with regards to undesired side effects such as eye itching, reddening, or tearing. Self-assessment questionnaires are usually filled in by study subjects in order to evaluate the level of discomfort, while redness and changes in its level are assessed by clinicians using the reference scales like the Efron scale or VBR. We believe that using an automated tool would increase the objectivity of such a study due to elimination of interobserver and intraobserver variability.

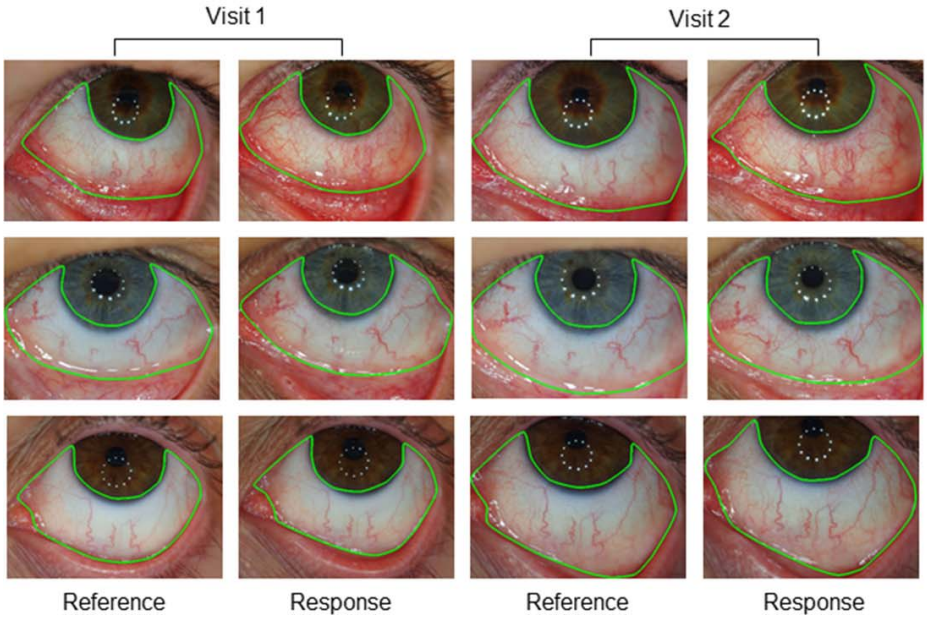
At the end of the last century, several researchers tried to objectivize ocular redness grading using photographic documentation. In 1990, Kjærgaard et al.[17] presented an experimental pipeline, in which five





Redness difference (response – reference)	Park	Amparo	Sárándi
Strong responder	0.004	0.129	0.127
Mild responder	0.006	0.067	0.064
No responder	-0.029	0.018	0.016

Figure 12. (A) Frequency distribution of the redness differences between the reference and response through three different redness scoring methods. (B) Illustrated example between a strong responder, mild responder, and a no responder and the values provided through the three different redness scoring methods.

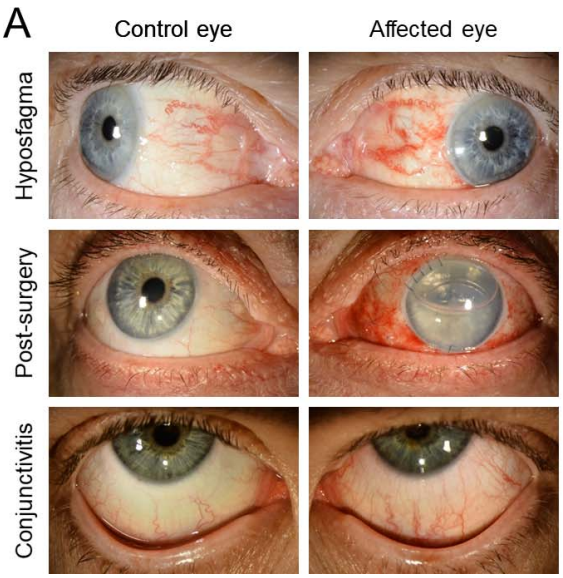


Subjective assessment	Redness difference (visit2- visit1) by human observer 3
Strong responder	0.99
Mild responder	0.75
No responder	0.11

Figure 13. Clinical application of the automated software by case examples in a conjunctival provocation test.

physicians used a descriptive scale in order to evaluate changes in ocular redness stimulated by the conjunctival provocation test. The final redness values were derived using statistics. The authors claimed a better sensitivity of their method as compared with traditional clinical observations. However, their method still is subjective, requires more resources (manpower), and does not support absolute measurements.

A further step toward objective quantification of ocular redness was the application of image processing to the photographic images. Such methods rely on machine-based quantification of integral redness of



B

Photo	Control eye			Affected eye		
	Park	Amparo	Sárándi	Park	Amparo	Sárándi
Hyposfagma	0.177 ± 0.023	0.218 ± 0.035	0.196 ± 0.035	0.263 ± 0.001	0.372 ± 0.003	0.352 ± 0.004
Post surgery	0.149 ± 0.014	0.341 ± 0.027	0.302 ± 0.028	0.335 ± 0.088	0.614 ± 0.078	0.588 ± 0.077
Conjunctivitis	0.142 ± 0.032	0.388 ± 0.013	0.330 ± 0.015	0.204 ± 0.028	0.435 ± 0.041	0.398 ± 0.041

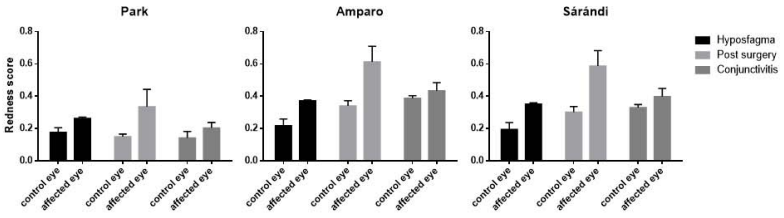


Figure 14. Three cases of ocular redness from the clinic. (A) A hyposphagma, postsurgical redness, and a mild form of conjunctivitis. (B) The table shows the numeric redness values of three pictures, averaged ± SD with visualization as bar graphs below the table.

the scleral region,[6,10,18–25] blood vessels dilation,[9,18,20–23,26–30] and degree of vascular branching,[31,32] or combination of these features. Integral redness is usually quantified as a ratio of pixels classified as red to the selected ROI[18,19,22,23] or as a result of arithmetical operations on color channels in different color models.[6,10,20,21,24,25] Blood vessels are usually segmented using edge detection,[9,21–23,25,26] thresholding with a prior enhancement,[9,18,20,28–30] or clustering[9] and are described in terms of percentage of vessel coverage,[9,18,20,21,23,25,27,29] vessel width,[20,27,30] relative redness of vessels,[20] and number of vessel segments.[20,30] Vascular branching is described using fractal analysis.[31,32]

Diseases and conditions may affect different regions of sclera[30]; it is beneficial to include in the ROI as much of sclera as possible. Fieguth and Simpson[21] postulated that automatic detection of sclera shall be straightforward, because its color is distinct from its surroundings. However, simple color thresholding fails in most of our images. The presence of shadows, light reflections, or excessively dilated blood vessels make it hard to distinguish between the sclera and surrounding regions. In contrast to the approaches using manual interaction for ROI detection[9,10,20,21,25,31] or color-based segmentation,[6] we therefore use texture information for automated sclera detection.

Sárándi et al.[6] proposed a fully automated scleral segmentation involving circular Hough transform[33] for iris subtraction and a combination of edge detection and thresholding in YUV color space for sclera localization. A common definition of a color space uses one luma component (Y′) and two chrominance components, called U and V. Their method works well if the sclera is evenly illuminated and highly distinguishable from the eyelid, but shadows or light reflections on the eyelid or the surrounding skin make the detection error prone. Furthermore, a high concentration of red blood vessels in the sclera often yields a segmentation failure.

It is still worth mentioning that according to visual inspection, there are outliers in our segmentation results that may undermine the stability of the general segmentation score. Erroneous ROI detections can be caused by a low quality of a photograph (nonsharp focus, uneven light, reflections) or by a similarity in textures. Blurred edges lead to loss of texture, which makes the detection of ROI and blood vessels not straightforward. The best way to deal with this problem is to control acquisition settings, that is, choosing the smallest aperture. In addition, we provided a customary tool for manual correction of the detected ROI, which still allows usage of images of lesser quality.

Another interesting observation was made with the respect to the provocation test: as it can be seen in Figure 10 for the response case, the redness in the second visit is lower than the redness in the first visit. We believe that this indicates that the provocation is better tolerated by the study subjects upon the second visit.

When we used clinical cases, the methods from Amparo et al.[10] and Sárándi et al.[6] provided a higher redness value for red eyes compared with the method of Park et al.[9] However, in all cases the methods showed higher signal for the affected eye. This indicates that using the contralateral eye as reference could be a proper solution when no follow-up visits are planned.

Almost all of the existing methods depend on a particular acquisition setup: all images shall be recorded with the same camera and illumination settings. However, this is not always possible, especially when comparing and analyzing a large amount of photographs taken in different laboratories (multicenter studies) or over various periods of time. Amparo et al.[10] introduced semiautomatic white balance correction using the Von Kries approach.[34] However, to our knowledge, full color normalization was not used before for ocular redness assessment. We will investigate this in the future.

In recent years, deep convolutional neural networks (CNN)[35] have gained their popularity in tasks of semantic image segmentation. Such techniques are able to classify the regions not only on a pixel level, but also on the object's shape as contextual information. Because the visible part of human sclera has a distinctive shape, we believe that it is possible to train such a classifier, which would enable recognition of human sclera with a considerably higher accuracy. We are planning to address this in our future work.

In summary, our study demonstrates that interactive user-guided segmentation leads to inconsistency in ocular redness scores driven by both intraobserver and interobserver variability. As an approach to this problem, automatic segmentation can be used. In the current study, we trained a simple random decision forest classifier, which in combination with an automatic ROI matching provided consistent results. Furthermore, our study has shown that the HSV color space resembling human color perception is better suited for redness scoring as it does not depend

on illumination and hand-crafted parameters. The outcomes of our proof of concept study are helpful for performing clinical trials targeted to assess ocular redness quantification over time.

Acknowledgments

The authors thank N. Kobelev and P. Schwehn for their assistance in implementing the software, and R. Mösges for sharing the conjunctival provocation test data set of Sárándi et al.[6] The research was conducted within the Chemelot InSciTe framework.

Disclosure: **E. Sirazitdinova**, None; **M. Gijs**, None; **C.J.F. Bertens**, None; **T.T.J.M. Berendschot**, None; **R.M.M.A. Nuijts**, None; **T.M. Deserno**, None

References

1. Cronau H, Kankanala RR, Mauger T. Diagnosis and management of red eye in primary care. *Am Fam Physician*. 2010;81:137–144.
2. McMonnies CW, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. *Am J Optom Physiol Op*. 1987;64:246–250.
3. Efron N. Clinical application of grading scales for contact lens complications. *Optician*. 1997;213:26–34.
4. Brien Holden Vision Institute (BHVI). Brien Holden Vision Institute grading scales. 2012; Available at: https://contactlensupdate.com/2012/11/20/brien-holden-vision-institute-grading-scales/. Accessed November 22, 2019.
5. Schulze MM, Jones DA, Simpson TL. The development of validated bulbar redness grading scales. *Optom Vis Sci*. 2007;84:976–983.
6. Sárándi I, Claßen DP, Astvatsatourov A, et al. Quantitative conjunctival provocation test for controlled clinical trials. *Methods Inf Med*. 2014;53:238–244.
7. Bertens CJF, Gijs M, van den Biggelaar FJHM, Nuijts RMMA. Topical drug delivery devices: a review. *Exp Eye Res*. 2018;168:149–160.
8. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int J Comp Vis*. 2004;60:91–110.
9. Park IK, Chun YS, Kim KG, Yang HK, Hwang JM. New clinical grading scales and objective measurement for conjunctival injection. *Invest Ophthalmol Vis Sci*. 2013;54:5249–5257.
10. Amparo F, Wang H, Emami-Naeini P, Karimian P, Dana R. The ocular redness index: a novel automated method for measuring ocular injection. *Invest Ophthalmol Vis Sci*. 2013;54:4821–4826.
11. Ho TK. Random decision forests. In: *Document Analysis and Recognition*. Montreal, Quebec, Canada: IEEE; 1995:278–282.
12. Witten IH, Frank E, Hall MA, Pal CJ. Data mining: practical machine learning tools and techniques. In: Green T, Pitts T, eds. *Data Management Systems*. 4th ed. Cambridge, MA: Morgan Kaufmann; 2017:242–243.
13. Arganda-Carreras I, Kaynig V, Rueden C, et al. Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification. *Bioinformatics*. 2017;33:2424–2426.
14. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
15. Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm ACM*. 1981;24:381–395.
16. Zuiderveld K. Contrast limited adaptive histogram equalization. In: Heckbert P, ed. *Graphics Gems IV*. San Diego, CA: Academic Press Professional, Inc.; 1994:474–485.
17. Kjærgaard SK, Pedersen OF, Taudorf E, Mølhave L. Assessment of changes in eye redness by a photographic method and the relation to sensory eye irritation. *Int Arch Occup Env Health*. 1990;62:133–137.
18. 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.
19. Horak F, Berger U, Menapace R, Schuster N. Quantification of conjunctival vascular reaction by digital imaging. *J Allerg Clin Immun*. 1996;98:495–500.
20. Papas EB. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci*. 2000;41:687–691.
21. Fieguth P, Simpson T. Automated measurement of bulbar redness. *Invest Ophthalmol Vis Sci*. 2002;43:340–347.
22. Wolffsohn JS, Purslow C. Clinical monitoring of ocular physiology using digital image analysis. *Cont Lens Anterior Eye*. 2003;26:27–35.
23. Peterson RC, Wolffsohn JS. Sensitivity and reliability of objective image analysis compared

to subjective grading of bulbar hyperaemia. *Br J Ophtalmol*. 2007;91:1464–1466.
24. Rodriguez JD, Johnston PR, Ousler GW III, Smith LM, Abelson MB. Automated grading system for evaluation of ocular redness associated with dry eye. *Clin Ophthalmol*. 2013;7:1197–1204.
25. Ferrari G, Rabiolo A, Bignami F, et al. Quantifying ocular surface inflammation and correlating it with inflammatory cell infiltration in vivo: a novel method. *Invest Ophthalmol Vis Sci*. 2015;56:7067–7075.
26. Villumsen J, Ringquist J, Alm A. Image analysis of conjunctival hyperemi. *Acta Ophthalmol*. 1991;69:536–539.
27. Guillon M, Shah D. Objective measurement of contact lens-induced conjunctival redness. *Optom Vis Sci*. 1996;73:595–605.
28. Owen CG, Fitzke FW, Woodward EG. A new computer assisted objective method for quantifying vascular changes of the bulbar conjunctivae. *Ophthalmic Physiol Opt*. 1996;16:430–437.
29. Dogan S, Astvatsatourov A, Deserno TM, et al. Objectifying the conjunctival provocation test: photography-based rating and digital analysis. *Int Arch Allerg Immun*. 2014;163:59–68.
30. Zhao WJ, Duan F, Li ZT, Yang HJ, Huang Q, Wu KL. Evaluation of regional bulbar redness using an image-based objective method. *Int J Ophtalmol*. 2014;7:71–76.
31. Schulze MM, Hutchings N, Simpson TL. The use of fractal analysis and photometry to estimate the accuracy of bulbar redness grading scales. *Invest Ophthalmol Vis Sci*. 2008;49:1398–1406.
32. Yoneda T, Sumi T, Takahashi A, Hoshikawa Y, Kobayashi M, Fukushima A. Automated hyperemia analysis software: reliability and reproducibility in healthy subjects. *Jpn J Ophthalmol*. 2012;56:1–7.
33. Lehmann TM, Kaupp A, Effert R, Meyer-Ebrecht D. Automatic strabometry by Hough-transformation and covarience-filtering. In: *Proceedings ICIP-94*. Los Alamitos, CA: IEEE; 1994:421–425.
34. Chong HY, Gortler SJ, Zickler T. *The von Kries hypothesis and a basis for color constancy*. In: *2007 IEEE 11th International Conference on Computer Vision (ICCV)*. Rio de Janeiro, Brazil: IEEE; 2007;1–8.
35. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436.