Image-based Eye Redness Using Standardized Ocular Surface Photography

Arno Lins, B.Sc

Abstract—**Systemic disease, inflammation, and irritation of the conjunctiva or sclera cause conjunctiva blood vessels to dilate, making the eye appear red [1]. The intensity of the redness is a crucial parameter in grading and monitoring ocular surface diseases. This ocular redness can be clinically graded by experts using various drawings- or photographic scales [2]. However, gradings based on these scales are subjective and suffer from inter-grader variability and bias. Hence, comparison and repeatability of different studies are difficult [2]. Objective methods based on digital ocular surface images were developed but are often limited by non-standardized imaging, i.e. variability in position, focus, illumination, and operator dependencies [1][3][4]. The Cornea Dome Lens (CDL) imaging system provides standardized ocular surface photography intending to overcome these limitations. Combined with a robust image analysis framework, there is the potential to make reproducible and accurate statements about eye redness. This work aims to develop the base for a reproducible pipeline to grade bulbar redness for the CDL imaging system. For this purpose, a semi-automated approach was established in the first step, extracting equal regions from a defined set of images. Subsequently, a Random Forest (RF) was trained for sclera segmentation. Finally, redness intensity based on Fieguth and Simpson [4] was tested as a feature to grade bulbar eye redness and further used to analyze a healthy set of volunteers. In addition, artificial eyes based on the Digital Bulbar Redness (DBR) scale were manufactured, imaged and analyzed for validation. The developed pipeline proved feasible to extract eye redness for the novel ocular imaging system for healthy eyes. Hence, it will serve as a solid baseline for further advancements, especially when adapting the redness extraction to more complex clinical cases with different ocular surface anomalies.**

Index Terms—**Ophthalmic photography · Ocular surface · Sclera · Classification · Feature extraction · Machine learning · Artificial eye.**

I. Introduction

A. Motivation

BULBAR ocular redness manifests due to the enlargement of conjunctival and episcleral blood vessels. The redness is induced either by inflammation and irritation of the bulbar conjunctiva or sclera or by a systemic disease [1][5]. Conjunctival hyperemia is one of the most common causes for visits to primary care physicians, optometrists, ophthalmologists, and emergency rooms [6]. It is a consistent sign of the ocular response to a pathologic stimulus. Moreover, it is a cardinal finding in a wide range of ocular surface disorders such as conjunctivitis, moderate-severe blepharitis, dry eye disease, and traumatic abrasions, among others [7]. Therefore, accurate analysis of the ocular surface redness is an important biomarker and the basis for further medical treatment planning.

A redness grading scale contains a range of images, each showing a different severity, from a white to a red eye. Experts like ophthalmologists generally grade an eyes redness by comparing the examined eye with those reference scales. Many different grading scales are available for this purpose in digital or in printed form. The reference images of a visual grading scale are either drawings (e.g. Efron) or photographs (e.g. CCLRU). Some of the most common scales are: Efron, (CCLRU) IER, VBR McMonnies/Chapman-Davis, Annunziato or Vistakon-Synoptik grading scale. Baudouin et al. analysed different scales and found scales cover different ranges of redness, and at least some scales are not linear across this range [2]. Due to this variety, Efron et al. recommend that clinicians stick with one scale for better and reproducible results. Unfortunately, this makes it challenging to compare scientific work which used different grading scales because the scales are not interchangeable with each other [8]. Hence, objective methods to grade redness are investigated in literature [9] [10] but often using different experimental setups, including different photographic devices, e.g. slit lamps and different operator dependant settings.

Binotti et al. [11] addressed that subjective scales are prone to variations of photographic devices, such as illumination, white balance, magnification or image resolution. The most common instrument to investigate bulbar redness is the slit lamp, a very powerful, versatile tool with many imaging modes to choose. However, a certain amount of know-how is required to operate the slit lamp. Furthermore, modern slit lamps have a camera integrated to acquire digital images and store them for documentation. Moreover, many different cameras are on the market. So even when experiments are reproducible, a comparison is difficult due to the wide variety of experimental setups.

Currently, only a few commercial alternatives exist. The Cornea Dome Lens (CDL) Imaging System aims to provide standardised ocular surface color photographs regarding position, lighting, focus and operator independence. With a novel lens design, high-resolution images of the cornea and the remaining anterior bulbar surface shall be possible.

B. Aim

This work aims to establish a reproducible and standardised pipeline to grade ocular redness with the CDL imaging system. The goal is to develop an automated way to get from raw images recorded by the device to a statement about the bulbar redness of the imaged eye. The different steps of the pipeline are investigated. This process demands image analysis through a pipeline of multiple steps to obtain top-level information

A. Lins is with the Department of Medical and Health Technologies, MCI, Innsbruck, Austria, e-mail: ar.lins@mci4me.at.

Manuscript received June 5, 2021; revised June 30, 2021.

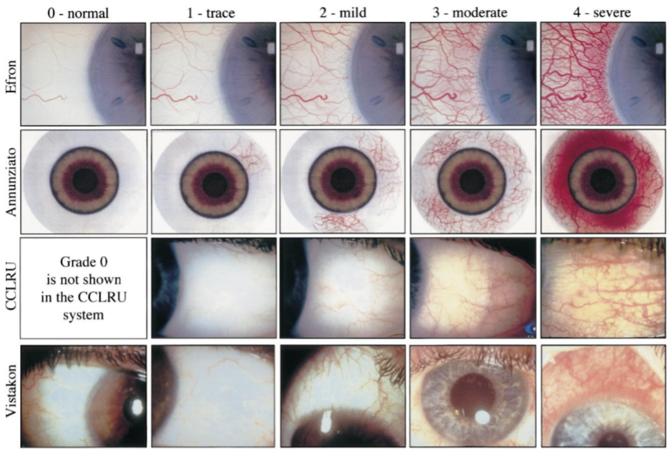


Fig. 1: Different subjective grading scales [8]: Annunziato and Efron representing drawings, Vistakon and CCLRU using images. The CCLRU additionally is divided into 4 grades.

about the ocular redness based on the high-resolution photos of the ocular surface. The major steps were: (i) finding a robust way to select sclera regions in the eye, (ii) defining a method to extract redness from the sclera regions. (iii) Furthermore, visualise bulbar redness based on the extracted redness. For the first step, data from healthy subjects were used to gain an impression of possibilities and limitations. For the second step, DBR-based data was used in addition to setting the feature methods' results in relation to gradings. In the last step, features were visualised on healthy data.

At the time of writing, the novel device is in its first-in-human trial. Hence no large clinical (pathological) dataset is available. Therefore, most of the work relies on healthy data, and it is unsure how well features and segmentation algorithms will work on pathological images. Therefore, this publication focus only on healthy data. Images of the DBR scale found in the literature are used as reference and validation of the feature's functionality to compensate for this.

Eye redness can have multiple reasons and occur due to enlargement of sclera or conjunctiva blood vessels [1][5]. This work does not distinguish between those types in terms of redness. Also, limbal redness can be treated separately but is not part of this work.

II. Methods

A. CDL Imaging Modality

The developed device provides a standardized ocular surface imaging system. A novel lens design enables high-resolution imaging of the visible ocular eye surface. The field-of-view is $21.3\,\mathrm{mm} \times 16.0\,\mathrm{mm}$ and the lateral resolution is approximately $15\,\mu\mathrm{m}$. The device includes a fixation target to minimize eye shivering and guarantee a centred view into the lens. Furthermore, software for eye tracking and an illumination unit is included. Chin rest and forehead band are used to stabilize the patient's head. The system aims for a standardized and operator-independent image acquisition. The aforementioned illumination unit shall provide an environmental independent image capturing. The recorded images have a full resolution of $4768\,\mathrm{px} \times 3580\,\mathrm{px}$ with a bit depth of $24\,\mathrm{bit}$ ($8\,\mathrm{bit}$ each of R-, G-, and B-channel) [5]. The device is not on the market yet and is currently in the first-in-human clinical evaluation.

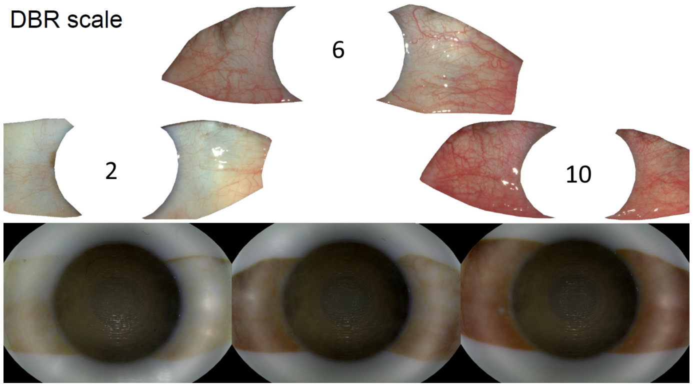


Fig. 2: From the DBR scale three different severity-grades were used as a template to generate artificial eyes

B. Healthy Dataset (CDL-2)

A dataset (CDL-2) was acquired with the CDL imaging system. This set includes 34 images of 17 healthy subjects with left and right eyes. Five out of 17 subjects were female. All other genders were male. Six subjects have ametropia, and all of them are male. The age of participants ranges from 26 to 46. Of all participants, 11 eyes are brown, four are green, and two are blue. Skin color range from two to six on the Fitzpatrick scale [12]. The ROI used for the dataset is $1000\,\mathrm{px} \times 1600\,\mathrm{px}$ placed centrally in vertical and $1600\,\mathrm{px}$ in horizontal direction from the iris center. Furthermore the ROIs are subdivided into 5×8 tiles containing sclera and non-sclera areas. Unlike the other sets, this one contains information about the type (sclera, non-sclera) of each tile. This information was added as a result of a labeling process (see Section II-F). The dataset is used to find an approach for automated sclera labelling and to review redness features which were extracted in the datasets mentioned before. The procedure was approved by the Research Committee for Scientific Ethical Questions from the UMIT TIROL (RCSEQ, 3012/22) and the MCI ethic commission (Kennzahl: 20220303). Informed consent was obtained from all volunteers prior to inclusion in this study.

C. DBR Phantoms

This small set contains three artificial eyes, which are 3D printed by an external company [13]. Those created eyes are using the DBR images as a template. Nasal and temporal images of grades 2, 6 and 10 are printed on three white eyes. In this way, three DBR phantoms were created. The data characteristics are again equal to those from the CDL-2 dataset regarding size and subdivisions. Hence, three temporal and three nasal ROIs graded as 2, 6 and 10 are available. This data was used to validate if the redness intensity feature can reproduce the grading to evaluate its usability.

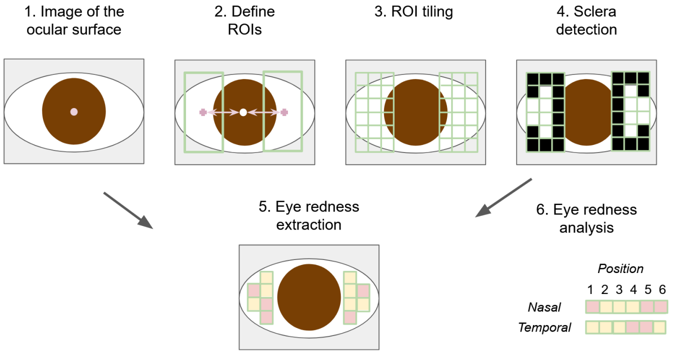


Fig. 3: Pipeline for eye redness extraction. From an initial ocular surface image (1) a ROI is defined (2) which is subdivided into tiles (3) and classified as sclera or non-sclera (4). Subsequently the ocular redness of the sclera tiles are extracted (5) analyzed and evaluated (6) [5].

D. Pipeline

The goal is to move from an initial imaged eye, like they are given by the CDL imaging system, towards a method which can describe the redness of this observed eye. The pipeline to reach this includes ROI definition, sclera segmentation, feature extraction, and feature evaluation. Once the investigated regions have been determined, the image's characteristics can be extracted to work as descriptors of eye redness. This is tested in the final "feature evaluation" step, which looks for a unique association with eye redness. Parts of the pipeline presented here were published within the scope of Ostheimer et al. [5]. Figure 3 visualizes this pipeline focusing on sclera segmentation and redness extraction.

E. Region of Interest

From all 34 images of the healthy subjects and the three DBR phantoms, a temporal and nasal region of the size 1000 px × 1600 px was extracted. The centre of both ROIs is placed 1600 px in horizontal direction left respectively right from the iris centre. The idea is to cover approximately 50% of the sclera and 50% of the non-sclera area to achieve a balanced dataset for machine learning when defining the problem for binary classification. This ROI was subdivided into 5 × 8 tiles with the size of 200 px × 200 px. The remaining tiles and features were used to train a random-forest classifier. The train-to-test ratio for the dataset was 80 to 20. The dataset was split with respect to the subjects to reduce the risk of overfitting because it was assumed that the correlation of tiles of the same eyes or subjects was higher than between different subjects.

F. Manual Labeling

The ROIs for the relevant data were further divided into 200 px × 200 px tiles leading to 2720 tiles for the CDL-2 dataset in total. Two observers labelled each tile with either sclera (i.e. pure sclera) or non-sclera (i.e. skin, iris, iris-sclera, skin-sclera). Tiles were dropped where the observer labels disagreed, leading to 2638 tiles. An annotation tool was used to label each tile efficiently [5]. The tool shows the actual processed image with two aforementioned 1000 px × 1600 px ROIs (Figure 4). An additional view is also provided containing a larger view of the current 200 px × 200 px tile, which has to be labelled by the operator. The operator can accept the tile as sclera or decline it by only using the keyboard. Afterwards, the program moves on to the next tile in coloum-major order. This was done with all images of the CDL-2 and DBR phantom dataset.

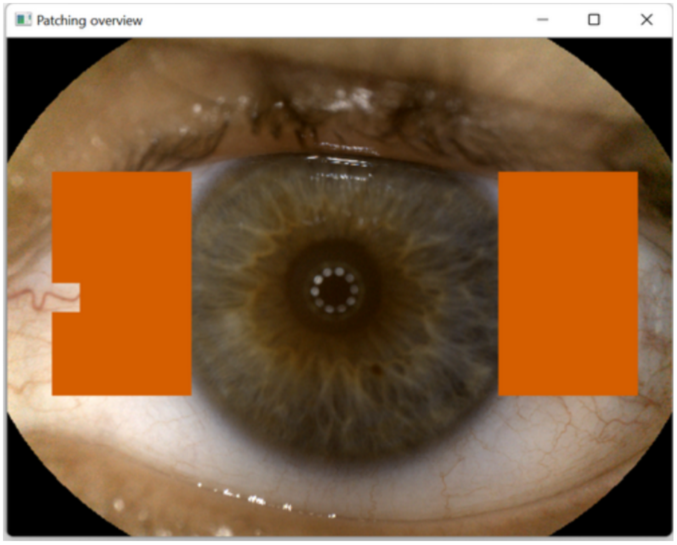


Fig. 4: Approaches for ROI-definition: Automated ROI based on iris fetching included in the imaging system. The orange rectangles define the ROI. The notch indicates the size of a tile.

G. Automated Segmentation

Segmentation was done on the tiles of the CDL-2 dataset using a random forest classifier. Therefore features were extracted from the tiles and fed to the classifier. To do so, gray-level co-occurrence matrices with 16 levels, 8 directions (equally distributed around 360°), and 4 distances (2,3,5,7) were calculated from each tile. From this, six haralick features provided by the scikit image library [14] (contrast (con), dissimilarity (dis), homogeneity ($homogen$), energy (E), correlation ($corr$) and angular second moment (ASM) were calculated. Additionally mean ($mean$) and standard deviation (std) for each color channel and the gray-scale tile was calculated. Furthermore, the Fieguth-redness was calculated and used as feature input, resulting in a list of 201 features. Three classifier were trained, one with all features, the second using only the ten most important features determined by the feature importance-function and the last based on feature correlation. The feature selection via correlation was based on an absolute correlation coefficient of 0.95 and dropping the feature with a lower correlation to the label. In this way, seven features were left.

The matrix created out of this CDL-2 dataset was split into train and test sets. First, the data was split by subjects to escape the threat of bias, which is created if tiles of similar images are in the train and test dataset. Since tiles with mismatching labels were discarded, the data was split into 76.5 % training and 23.5 % test data, respectively.

H. Feature extraction

Two intensity-based features were selected from the literature to test them on the CDL-1 and DBR data. Namely, the approach by Fieguth and Simpson [4]. The approach tries to make a statement about eye redness by investigating relations between the image color channels. The Fieguth-redness [4] is defined as:

$$f_r(S) = \frac{1}{\|S\|} \sum_{i \in S} \frac{2(S_R)_i - (S_G)_i - (S_B)_i}{2[(S_R)_i + (S_G)_i + (S_B)_i]} \qquad (1)$$

S is the investigated image segment, $S_{R,G,B}$ the red, green or blue image channel of the image segment and $|S|$ is the denominator representing the total number of pixels and works as a normalization. By the definition of the function values are limited between $f_r(S) = -0.5$ and $f_r(S) = 1$. Black pixels in the image should be avoided as they result in a division by zero. The theoretical possible maximum would initiated by a red image, $f_r(S) = 0$ by a white image and the minimum by a blue or green image. Additional blurring with a kernel of size $7\,\text{px} \times 7\,\text{px}$ and a $\sigma = 1$ was applied beforehand to reduce the effect of noise.

I. Feature Evaluation

The idea with the artificial eyes (see Fig. 2) is that with additional knowledge about the severity-grade, measurable differences in the output of the feature are expected. Because the artificial eyes are based on the DBR scale, images extracted with the novel system can validate the device as it enables to set feature-values in relation to an already established grading scale.

Feature values were extracted on the ROIs of the DBR phantoms and the CDL-2 dataset and compared quantitatively with each other. The phantom dataset includes severity labels, and from the CDL-2 dataset it is assumed that all subjects are healthy. Therefore values from the CDL-2 images are expected at the constant (lower) level compared to those of the DBR images.

III. Results

A. ROI Definition for Automation

To define a fixed ROI for high-throughput image processing of the CDL-2 dataset a ROI was defined based on two requirements: (i) Consider size and position where sclera area is present in most images and (ii) ROI should also contain non-sclera regions aiming for a $50/50$ ratio. Following this the region was set to $1600\,\text{px} \times 1000\,\text{px}$ with a distance of $1600\,\text{px}$ to the iris center measured at the center of the region (Figure 4). This applies to both nasal and temporal regions.

model	Features (descending order of importance for RF_{10})
RF_{10}	$mean_G, corr_{(7,315)}, corr_{(7,45)} mean_R,$ $corr_{(7,225)}, std_{gray}, mean_B,$ $std_G, corr_{(5,45)}, corr_{(5,0)}$
$RF_{(corr)}$	$homogen_{(5,90)}, homogen_{(5,180)},$ $E_{(2,45)}, corr_{(7,90)}, mean_R, std_R, f_r$

TABLE I: Most relevant features for the RF classifier by feature importance and correlation.

B. ROI Tiling for Sclera Segmentation

As the data is used for sclera segmentation, the regions are divided into small tiles that the classifier must determine. The tiles had to be small enough so that only tiles containing sclera could be found (because only then does the tile get labelled as such (see Section II-F)), but large enough to detect vessels and other structures. Therefore, and with the additional information from Section III-A the tile size was set to $200\,\text{px} \times 200\,\text{px}$ resulting in 8×5 tile regions.

C. Classification of Sclera/Non-sclera

1) Manual Classification: In a joint work, [5] the CDL-2 dataset with 2720 non-overlapping tiles was labelled by two observers. The findings presented in Ostheimer et al. are an agreement between those two observers of 96.99%, which means a disagreement of 82 tiles. With respect to this, 47.57% of the tiles were labeled as sclera and 52.43% as non-sclera tiles. In Figure 5 the agreement between both observers is indicated as non-sclera (black), sclera (white) and disagreement (gray).

2) ML Classification: As described in Section II-G, three RF classifiers were trained, each with a different feature set. One set included all features, the remaining two only a subset which can be found in Table I. The reduction was done by feature importance [15] and correlation. The last one resulted in $n = 7$ features. The split between test and train data was done random but with respect to subject dependency to ensure images of eyes from the same subject were included in only one of the two datasets to reduce the risk of bias. The RF classifier trained with its associated part of the CDL-2 dataset achieved on the additional test dataset and with the features from the correlation-approach the best results. The accuracy was 0.974, precision and F1-score were 0.97, and recall 0.98 (Table II). Investigating images on which the classifier predicted its label showed difficulties for the classifier to distinguish between birthmarks and non-sclera regions, as one can see from Figure 5 (A) tile (A4). Also, one can see by comparing tile (A3) from Figure 5 (A) less conservative labelling by the classifier than one of the observers.

model	n	accuracy	precision	recall	f1-score
RF_{all}	201	0.971	**0.97**	0.97	**0.97**
RF_{10}	10	0.943	0.94	0.94	0.94
RF_{corr}	7	**0.974**	**0.97**	**0.98**	**0.97**

TABLE II: Performance values of the random forest classifier to label tiles as sclera.

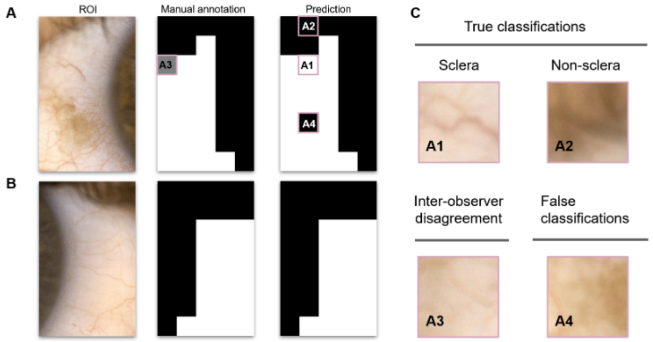


Fig. 5: (A) Sclera segmentation for a ROI with disagreement between two observers (A3) and between observers and RF prediction (A4). (C) shows those tiles in detail. (B) Sclera segmentation for a ROI with no disagreement between either observer nor RF prediction.

D. Artificial Eyes

Investigating the artificial eyes based on the DBR scale ($P = phantoms$), 78 tiles were labelled as sclera for the nasal and 91 for the temporal side. For each side (nasal and temporal) a linear fit was applied for the sclera-tiles. Also, the mean values of each ROI can be found in Figure 6a. The slope nasal and temporal is $k_n^P = 0.0018$ and $k_t^P = 0.0023$. From the graph of the original publication [16] in Figure 6b a slope of approximately $k_{n,t}^{DBR} = 0.0025$ is derived from both sides. All values for the artificial eyes differ by an offset of approximately $d = 0.35$ from the original values [16].

E. Redness Extraction

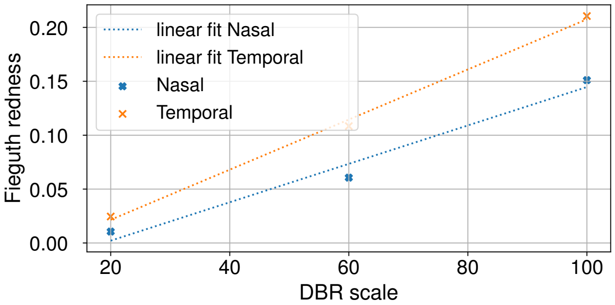
From Figure 7 one can see the redness values for the healthy subjects f_r^{CDL-2} stay on a lower level μ_r^{CDL-2} with respect to the DBR data f_r^{DBR} and do not spread across the range of the DBR values ($f_{r_{min}}^{DBR}$, $f_{r_{max}}^{DBR}$). By reviewing the data, outliers for $eye = 2$ and $eye = 16$ could be identified as Birthmarks (e.g. Figure 9b). From the DBR phantoms, a rise for the different grades is visible.

F. Subject Characteristics

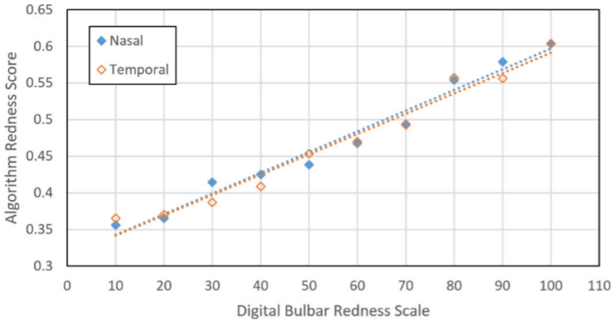
Redness intensity was applied to the CDL-2 dataset and investigated regarding subject characteristics. Scatter and boxplots are shown in Figure 8. For the characteristics of sex, eye color, and skin color all median values, as well as the interquartile range, are on a similar level. Higher values for individual variables of a certain characteristic do exist but are outliers, e.g. $f_r \approx 0.3$ is from a birthmark. No trend can be recognized by a scatter plot of redness versus age. Two outliers can be recognized for eye 2 and eye 16.

G. Visualization

A heatmap was created to visualize the extracted redness, For each tile of a ROI labelled as sclera, the Fieguth-redness was calculated and visualized in a corresponding red tone. Non-sclera tiles were masked black. The limits were set approximately according to the highest and lowest value of the DBR phantoms which were $f_{r_{max}}^P = 0.5$ and $f_{r_{min}}^P = 0$, respectively. This range also covers all values from other datasets (Figure 7). For better interpretation, the heatmap was subsequently placed transparent on top of the original image (Figure 9). From this visualization, one can see which parts of the sclera are taken into account and what sub-regions are suggested by the algorithm as more reddish.



(a) Mean redness of the DBR phantoms



(b) Redness of DBR reference images

Fig. 6: (a) Upper plot visualizes the mean redness value of each artificial eye and ROI including a linear fit to compare with (b) the results from the DBR scale from literature[16].

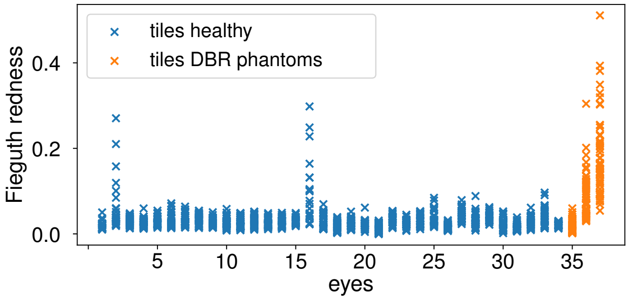


Fig. 7: Fieguth-redness for all tiles of all healthy eyes (CDL-2) and DBR phantoms. Constant lower values for healthy data and higher values for rising severity in the DBR phantom data.

IV. Discussion

A. Region of Interest

Sclera segmentation is necessary to adjust the region of interest to the patient's eyelid opening to ensure accurate

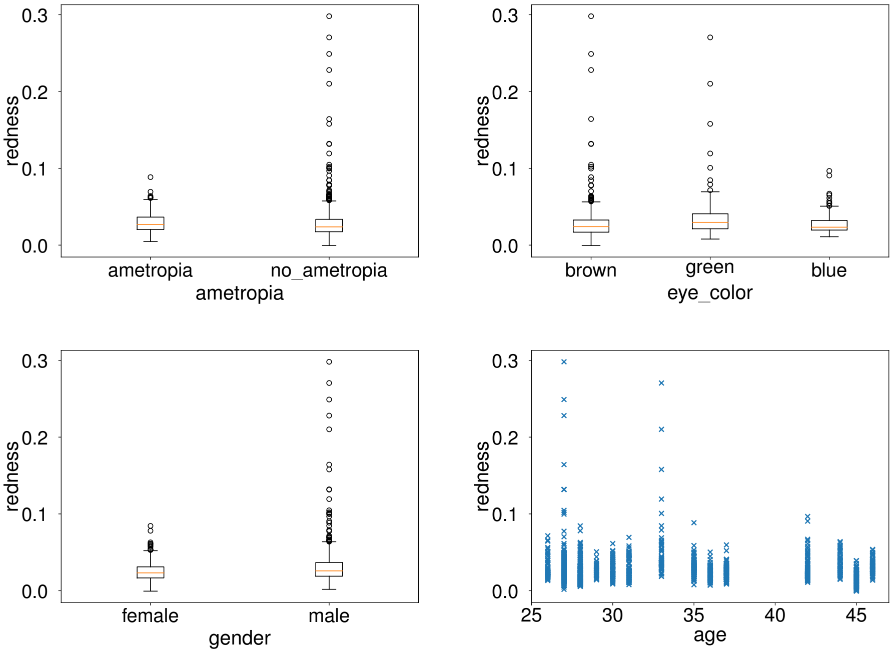


Fig. 8: Influence of different subject characteristics to Fieguth redness.

redness detection. The fixed ROI was divided into tiles, which were consequently manually annotated, and carried out by two observers.

The manual segmentation reached a good amount of agreement among the observer probably because both observers discussed and agreed on the requirements necessary to label a tile as sclera or non-sclera. The reason why there were still disagreements was perhaps due to two factors.

First, the inconsistency of the upper tiles is most likely due to blurred eyelids in the image. Second, the lack of agreement at the iris edge is probably due to the color gradient around the iris that results from the anatomical structure. This makes it difficult to decide where the iris begins, and the sclera ends. Including a reliable iris and limbus segmentation might overcome this problem in the future.

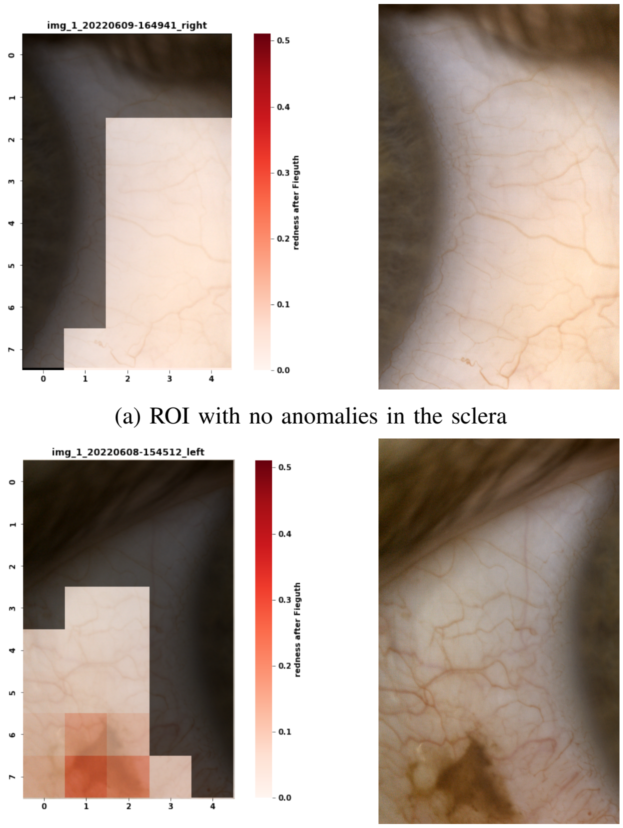
The manual labels were furthermore used to train an ML algorithm to see how well this step can be automated. The performance from the RF seems to be a promising start, but the images used to train the random forest classifier were all from healthy subjects. In the future, detecting and segmenting the sclera of pathological eyes will be necessary. Therefore, the performance results should be taken with caution. Furthermore, it is unclear how well the classifier performs in pathological eyes or whether similar performance results can be expected when additional pathological images are used as training data.

Another limitation is that the amount of the dataset is relatively small. The regions were split into tiles where we assume a certain correlation from tiles of the same ROI. Hence, to train the classifier, the data was split according to patients to provide an appropriate insight into the model's generalization. Because of the small dataset and the splitting mentioned above, the train- respectively test-dataset is less diverse. This means, for example, no blue eyes might be in one dataset, leading to weaker performance. However, our classifier's results seem good because they reach a performance comparable to the inter-observer agreement of the manual classification. This might change if we not only want to distinguish between sclera and non-sclera but also find iris, eye leashes and anomalies like birthmarks or lesions. Especially birthmarks or other lesions might be interesting to distinguish. As it can be seen from Figure 9b and Figure 5, it might wrongly be labelled as non-sclera, or if not, the algorithm rate the birthmark as very red.

Comparable approaches try to segment the whole sclera in an image, such as Sirazitdinova et al. [3], Sárándi et al. [17] or Sánchez et al [18]. This is difficult, and requires multiple steps and a certain amount of computational effort. Other approaches [19] [20] are based on manual segmentation by marking a rectangle or drawing a spline. Therefore, the approach presented here is a good alternative as it has low computational effort but still tries to recognize the individual area size. The approach could also be expanded to larger regions. Furthermore, the optimal tile size should be the subject of future investigations.

Feature extraction for the random forest classifier was applied using all six Haralick features the sci-kit image library provides. Expanding the feature-set could be considered when more than a binary classification is intended. Additionally, each colour channel's mean and standard deviation in the RGB color space was used. Other color spaces could be considered in the future. [17][21].

Extension of the approach to Deep Learning would circumvent the problem of finding the right features for optimal sclera segmentation. Moreover, other methods may be more suitable in the future, especially if the problem complexity increases by including other classes such as iris, eye leashes or anomalies



(a) ROI with no anomalies in the sclera

(b) ROI with anomalies in the sclera

Fig. 9: Example of visualization of extracted features on healthy ROI with no anomalies (a) and with a birthmark (b). Fieguth-redness next to the extracted ROI and the impact to the grading of anomalies in the sclera.

like birthmarks or lesions.

B. Feature Extraction

An advantage of intensity-based methods is the scale invariant which is especially useful when comparing datasets with different shapes and sizes. Fieguth and Simpson [4] mentions that their practical values move approximately in the range of $0 \leq f_r^{lit} \leq 0.25$. Macchi et al. [16] which invented the DBR scale and which algorithm for the scale is based on [4] reached values between $0.35 \leq f_r^{scale} \leq 0.6$. In both cases, the distance from most to least red is around 0.25. In our case, three artificial eyes based on the DBR scale were imaged and analyzed individually, nasal and temporal. The printed eyes represent grades 2, 6 and 10 on the DBR scale. The extended range is around 0.16 and 0.20 for the nasal and temporal side of the DBR phantoms, respectively. If the same images were used, one had to assume to get similar results. Results clearly show an offset ($d = 0.35$) but the slopes seem to be comparable ($k_t^P = 0.0023, k_n^P = 0.0018, k_{n,t}^{DBR} = 0.0025$). The range between most and least redness is also close to the values known from the literature [4][16]. However, these results also validate the ability of the device in combination with the algorithm to detect redness changes.

C. Subject Characteristics

McMoonies and Ho, and Murphy et al. found differences in redness in gender and age[22][23]. Murphy et al. found 0.2 units higher for males than females on an interpolated CCLRU scale ranging from one to four [23]. Furthermore, the authors found an increase in redness by 0.05 units per decade. Our subjects age in the CDL-2 data range from mid-twenties to mid-forties. The mentioned study has participants from the age of 16 to 77. The data studied here (CDL-2) show no trend in age, sex, or other patient characteristics. However, larger studies with healthy subjects, as mentioned above, show such trends. Therefore, it would be interesting to collect a larger dataset to detect further redness population characteristics.

D. Visualization

The upper and lower limits for the color scale should be considered for the visualization of the redness intensity. Choosing theoretical limits might result in unnecessary low sensitivity. Fieguth and Simpson [4] mentions values between $0 \leq f_r \leq 0.25$ for these experiment. A threshold value of $f_r = 1$ would result in all values having similar colors. On the other hand, if the threshold value were set too low, moderately red eyes could not be distinguished from strongly red eyes—the same hold for the lower limit. Hence, the actual parameters should be set after evaluating more clinical data. Currently, chosen values are 0 and 0.5.

Proper visualization is essential to show and guide the operator to interesting regions; additionally, areas which are not covered by the algorithm mustn't be hidden from the operator as they might include important information. Furthermore, this becomes more important when the pipeline is integrated into decision support software.

V. Conclusion

The novel ocular surface imaging device has the potential to objectively and reproducible grade bulbar eye redness based on its recorded images. The first steps in this direction were taken with the procedure presented here and its results. An ML approach showed promising results and will enable automated sclera segmentation of new unseen images. For a simple sclera vs non-sclera classification in healthy subjects, a RF classifier seems sufficient, neglecting the influence of birthmarks etc. Deep learning approaches might be necessary if more classes like birthmarks, iris etc. should be predicted because those make accurate segmentation complex and manual feature engineering cumbersome.

The severity of the DBR phantom images showed a good correlation with the intensity feature. This supports the assumption that the intensity feature is a working measure for bulbar redness on the novel imaging system.

No differences in subject characteristics could be detected from a small dataset and with qualitative measures. A detailed statistical analysis was not performed as the dataset was considered too small.

Visualization of the redness extraction will be an important task to handle as it is the interface to clinicians. In addition, it forms the basis for whether important information is recognized more quickly or is overlooked. Therefore future efforts should address this issue by evaluating and taking a closer look at the clinical data.

Acknowledgment

The author would like to thank Marco Augustin form OCCYO GmbH for the supervising, Philipp Ostheimer, Phd Student at UMIT for the collaboration and Benjamin Massow for the MCI internal support.

References

[1] H. Pult, P. J. Murphy, C. Purslow, J. Nyman, and R. L. Woods, "Limbal and bulbar hyperaemia in normal eyes." *Ophthalmic & Physiological Optics*, vol. 28, no. 1, pp. 13–20, jan 2008. [Online]. Available: http://dx.doi.org/10.1111/j.1475-1313.2007.00534.x

[2] C. Baudouin, K. Barton, M. Cucherat, and C. Traverso, "The measurement of bulbar hyperemia: challenges and pitfalls." *European Journal of Ophthalmology*, vol. 25, no. 4, pp. 273–279, aug 2015. [Online]. Available: http://dx.doi.org/10.5301/ejo.5000626

[3] E. Sirazitdinova, M. Gijs, C. J. F. Bertens, T. T. J. M. Berendschot, R. M. M. A. Nuijts, and T. M. Deserno, "Validation of computerized quantification of ocular redness." *Translational vision science & technology*, vol. 8, no. 6, p. 31, nov 2019. [Online]. Available: http://dx.doi.org/10.1167/tvst.8.6.31

[4] P. Fieguth and T. Simpson, "Automated measurement of bulbar redness." *Investigative Ophthalmology & Visual Science*, vol. 43, no. 2, pp. 340–347, feb 2002. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11818375

[5] P. Ostheimer, A. Lins, B. Massow, B. Steger, D. Baumgarten, and M. Augustin, *Extraction of eye redness for standardized ocular surface photography*. MICCAI.

[6] R. B. Singh, L. Liu, A. Yung, S. Anchouche, S. K. Mittal, T. Blanco, T. H. Dohlman, J. Yin, and R. Dana, "Ocular redness - II: Progress in development of therapeutics for the management of conjunctival hyperemia." *The ocular surface*, vol. 21, pp. 66–77, jul 2021. [Online]. Available: http://dx.doi.org/10.1016/j.jtos.2021.05.004

[7] F. Pérez-Bartolomé, C. Sanz-Pozo, J. M. Martínez-de la Casa, P. Arriola-Villalobos, C. Fernández-Pérez, and J. García-Feijoó, "Assessment of ocular redness measurements obtained with keratograph 5M and correlation with subjective grading scales." *Journal Francais d'Ophtalmologie*, vol. 41, no. 9, pp. 836–846, nov 2018. [Online]. Available: http://dx.doi.org/10.1016/j.jfo.2018.03.007

[8] N. Efron, P. B. Morgan, and S. S. Katsara, "Validation of grading scales for contact lens complications." *Ophthalmic & Physiological Optics*, vol. 21, no. 1, pp. 17–29, jan 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11220037

[9] W.-J. Zhao, F. Duan, Z.-T. Li, H.-J. Yang, Q. Huang, and K.-L. Wu, "Evaluation of regional bulbar redness using an image-based objective method." *International journal of ophthalmology*, vol. 7, no. 1, pp. 71–76, feb 2014. [Online]. Available: http://dx.doi.org/10.3980/j.issn.2222-3959.2014.01.13

[10] I. K. Park, Y. S. Chun, K. G. Kim, H. K. Yang, and J.-M. Hwang, "New clinical grading scales and objective measurement for conjunctival injection." *Investigative Ophthalmology & Visual Science*, vol. 54, no. 8, pp. 5249–5257, aug 2013. [Online]. Available: http://dx.doi.org/10.1167/iovs.12-10678

[11] W. W. Binotti, B. Bayraktutar, M. C. Ozmen, S. M. Cox, and P. Hamrah, "A review of imaging biomarkers of the ocular surface." *Eye & contact lens*, vol. 46 Suppl 2, pp. S84–S105, mar 2020. [Online]. Available: http://dx.doi.org/10.1097/ICL.0000000000000684

[12] T. B. Fitzpatrick, "Soleil et peau," *Journal de Médecine Esthétique*, no. 2, p. 33–34, 1975.

[13] [Online]. Available: https://www.eyecre.at/

[14] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and scikit-image contributors, "scikit-image: image processing in python." *PeerJ*, vol. 2, p. e453, jun 2014. [Online]. Available: http://dx.doi.org/10.7717/peerj.453

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[16] I. Macchi, V. Y. Bunya, M. Massaro-Giordano, R. A. Stone, M. G. Maguire, Y. Zheng, M. Chen, J. Gee, E. Smith, and E. Daniel, "A new scale for the assessment of conjunctival bulbar redness." *The ocular surface*, vol. 16, no. 4, pp. 436–440, oct 2018. [Online]. Available: http://dx.doi.org/10.1016/j.jtos.2018.06.003

[17] I. Sárándi, D. P. Cla, A. Astvatsatourov, O. Pfaar, L. Klimek, R. Mösges, and T. M. Deserno, "Quantitative conjunctival provocation test for controlled clinical trials." *Methods of Information in Medicine*, vol. 53, no. 4, pp. 238–244, jun 2014. [Online]. Available: http://dx.doi.org/10.3414/ME13-12-0142

[18] M. L. Sánchez Brea, N. Barreira Rodríguez, A. Mosquera González, K. Evans, and H. Pena-Verdeal, "Defining the optimal region of interest for hyperemia grading in the bulbar conjunctiva." *Computational and mathematical methods in medicine*, vol. 2016, p. 3695014, dec 2016. [Online]. Available: http://dx.doi.org/10.1155/2016/3695014

[19] F. Amparo, H. Wang, P. Emami-Naeini, P. Karimian, and R. Dana, "The ocular redness index: a novel automated method for measuring ocular injection." *Investigative Ophthalmology & Visual Science*, vol. 54, no. 7, pp. 4821–4826, jul 2013. [Online]. Available: http://dx.doi.org/10.1167/iovs.13-12217

[20] T. Yoneda, T. Sumi, A. Takahashi, Y. Hoshikawa, M. Kobayashi, and A. Fukushima, "Automated hyperemia analysis software: reliability and reproducibility in healthy subjects." *Japanese Journal of Ophthalmology*, vol. 56, no. 1, pp. 1–7, jan 2012. [Online]. Available: http://dx.doi.org/10.1007/s10384-011-0107-2

[21] N. Curti, E. Giampieri, F. Guaraldi, F. Bernabei, L. Cercenelli, G. Castellani, P. Versura, and E. Marcelli, "A fully automated pipeline for a robust conjunctival hyperemia estimation," *Applied Sciences*, vol. 11, no. 7, p. 2978, mar 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/7/2978

[22] C. W. McMonnies and A. Ho, "Conjunctival hyperaemia in non-contact lens wearers," *Acta Ophthalmol (Copenh)*, vol. 69, no. 6, pp. 799–801, Dec 1991.

[23] P. J. Murphy, J. S. C. Lau, M. M. L. Sim, and R. L. Woods, "How red is a white eye? clinical grading of normal conjunctival hyperaemia." *Eye*, vol. 21, no. 5, pp. 633–638, may 2007. [Online]. Available: http://dx.doi.org/10.1038/sj.eye.6702295

Arno Lins is a student at the Department of Medical and Health Technologies, MCI, Innsbruck, Austria.