Assessment of the repeatability in an automatic methodology for hyperemia grading in the bulbar conjunctiva

Luisa Sánchez Brea*, Noelia Barreira Rodríguez*, Antonio Mosquera González† and Katharine Evans‡
*Department of Computer Science, University of A Coruña, Spain, Email: luisa.brea@udc.es
†Department of Electronics and Computer Science, University of Santiago de Compostela, Spain
‡School of Optometry and Vision Sciences, Cardiff University, Wales

***Abstract*—When the vessels of the bulbar conjunctiva get congested with blood, a characteristic red hue appears in the area. This symptom is known as hyperemia, and can be an early indicator of certain pathologies. Therefore, a prompt diagnosis is desirable in order to minimize both medical and economic repercussions. A fully automatic methodology for hyperemia grading in the bulbar conjunctiva was developed, by means of image processing and machine learning techniques. As there is a wide range of illumination, contrast, and focus issues in the images that specialists use to perform the grading, a repeatability analysis is necessary. Thus, the validation of each step of the methodology was performed, analyzing how variations in the images are translated to the results, and comparing them to the optometrist's measurements. Our results prove the robustness of our methodology to various conditions. Moreover, the differences in the automatic outputs are similar to the optometrist's ones.**

I. Introduction

Hyperemia is the occurrence of vessel engorgement in a certain tissue. Specifically, this article, focuses on the hyperemia level on the bulbar conjunctiva. In this area, hyperemia can happen as a consequence of normal bodily processes, but also as a symptom of certain pathologies, such as allergic conjunctivitis [1] or dry eye syndrome [2]. These pathologies have a high incidence among the world population, and their prompt diagnosis is desirable from both the medical and economical points of view.

The manual process that specialists have to tackle in order to evaluate the hyperemia level is time-consuming. Moreover, the grading is not objective nor repeatable, showing high levels of both inter- and intra-expert subjectivity [3]. The process starts by filming a video or capturing a picture of the patient eye. In the case of the video, the specialists have to search through it in order to find the frame that offers the best depiction of the conjunctiva. Then, they analyze that frame, looking for indicators of hyperemia, such as general redness of the conjunctiva or quantity of vessels. Finally, the optometrists compare the patient's eye with a grading scale. Grading scales are collections of images that depict levels of severity. One of the most widely used scales is the Efron grading scale, which is shown in Fig. 1.

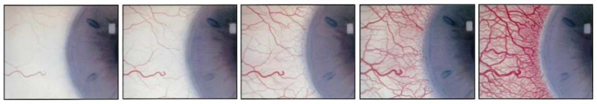

Fig. 1. Efron grading scale for bulbar hyperemia. It comprises five drawings, from 0 (left, lowest level) to 4 (right, highest level).

The automation of this procedure requires four steps. The first step is the selection of the best frame of the video sequence. The second step, the segmentation of the region of interest, comprising most of the conjunctiva. The third step is the computation of several image features within the region of interest, such as level of red in several color spaces or vessel quantity. Finally, the fourth and last step combines the values of these features in order to return a grade in the given scale.

There are few attempts on automatic hyperemia grading in the literature, and they either are not fully automatic or have a different aim than the emulation of the specialists' evaluation. Thus, in [4] an automatic method for the evaluation of dry eye redness based on image processing techniques is proposed. The authors compute two image features (redness intensity and the prominence of horizontal vessels) and analyze their concordance with the optometrists' gradings. The region of interest is segmented manually, and the images belong to patients with dry eye syndrome. In [5], a rectangular region of interest was manually selected in order to compute the pixel coverage of that part of the conjunctiva. However, the main focus of this work is the reliability of the process instead of the grading results. The images in these works have been captured in the same conditions and, therefore, they seem to present an homogeneous illumination. Regarding the definition of image features that are related to hyperemia, several works have proposed objective metrics through the years [6]–[8]. To the best of our knowledge, there are no other works that propose a fully automatic approach to bulbar hyperemia evaluation. Specifically, a key aspect such as the study of image features and their combination is missing in the literature.

An automatic methodology for hyperemia grading was developed in order to mimic the specialists' process while ensuring objectivity, repeatability, and an improvement of the invested time [9]. Image processing algorithms are applied to the input image in order to segment the region of interest. Then, several image features are computed in this region, based on measuring the intensity level of a certain hue and

978-1-5090-6182-2/17/$31.00 ©2017 IEEE

in the disposition of the vessels, which are located by an edge detection method. Finally, the image features are transformed to the value in the grading scale by means of regression techniques. During the development of this methodology, a recurrent drawback appeared: the inputs of the system (videos or images of the patient eye) present a high variability regarding conditions such as illumination and focus. Besides, in some cases, such as wearing contacts, the grading should not vary, but the image characteristics are different, as depicted in Fig. 2. The presence of contacts can affect to the vessel counting features, since the lens edges can get mixed up with the vessel edges.

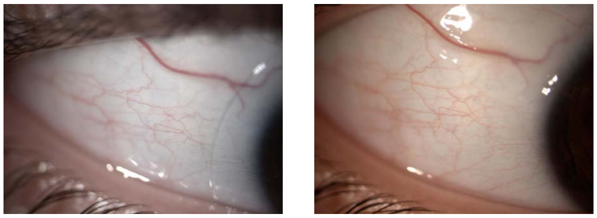

Fig. 2. Two images of the same eye with and without contacts. It can be observed how one of the images was took under a brighter light, which can affect the color based features of the image.

Therefore, in this work an exhaustive repeatability study is performed. The expert's gradings are studied in order to establish which level of discrepancy is to be expected between images of the same eye in slightly different circumstances. Our objective is to study the influence of several conditions in the subjective grading, and how they affect each step of the methodology. To that end, the segmentation of the region of interest is evaluated. Next, the results of the automatic methodology are validated in a large image set. Then, the impact that the different variations in the image's conditions have in the results and in each stage of the automatic methodology is analyzed. Finally, the differences between the expert's evaluation and the automatic approach in the same image are also analyzed. This knowledge will allow us to verify the robustness of the algorithms, and to further compare the objective and subjective approaches.

This article is structured as follows: Section II will provide a brief overview on the image set and the automatic methodology, and will explain in detail the repeatability tests performed. Section III will show the obtained results. Finally, Section IV will depict the conclusions and future lines of work.

II. Methodology

A. Dataset

The database used in this work was obtained as part of an study regarding contact lenses comfort, conducted at the School of Optometry and Vision Sciences from the Cardiff University[1]. 35 participants took part in the study, that consisted of four checkups. The first one (baseline) consist of pictures of the patients' eyes without contacts. The second checkup took place two weeks after the first one, and consists of images of the eyes while wearing contacts. The patients were asked to wear contacts each day during this two week period. Then, the third checkup took place after a 7-day washout period (non-wearing contacts period), and depicts the patients' eyes without contacts. Finally, the fourth checkup took place after another two week trial of contacts, and again depicts the patients while wearing contacts.

During each checkup, 4 types of images were taken, depicting both eyes and both sides of the eye. The four types are left eye, nasal side (*LEN*); left eye, temporal side (*LET*); right eye, nasal side (*REN*); and right eye, temporal side (*RET*). An example of these 4 different combinations of eye and side is depicted in Fig. 3. All of them show a side view of the eye and, therefore, they have a similar disposition. In some cases, several pictures of a particular type were taken, this is, there may be several images depicting the same patient, checkup, eye, and side. In these cases, all the images of the same type were processed by the automatic methodology and the results were averaged. Regarding the manual grading, only one hyperemia evaluation was performed by a single optometrist for each type in each checkup of each patient. However, it must be noted that the objective of this work is not to study the inter-expert variability, which has been the focus of a previous work [3].

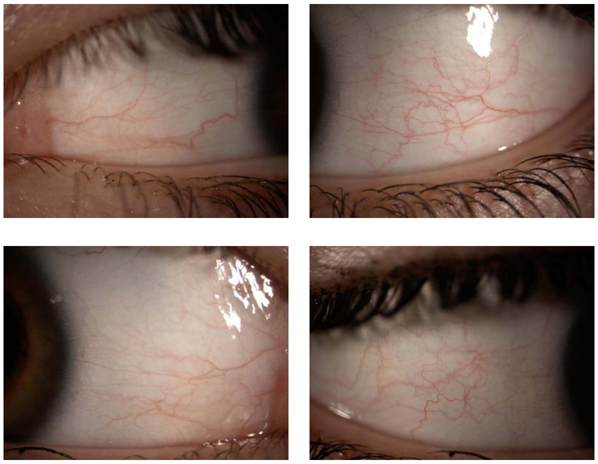

Fig. 3. Different eyes and sides for a certain patient and checkup. From left to right and top to bottom: LEN, LET, REN and RET.

As our objective is to assess the effect that certain alterations have in the methodology, two image subsets were created, one for each alteration that is going to be analyzed, that consist of 10 pairs of images each. Each pair depicts two images of the same patient, same eye, and same side. One of the images, labeled as *reference*, presents optimal conditions for performing hyperemia grading, while the second one, labeled as *altered* presents the analyzed alteration. The first subset, S_{cont}, includes images with and without contact lenses. The second subset, S_{clean}, depicts the eyes with and without remains of a cleaning lotion. For this second test, both images belong to the same checkup and, therefore, were taken minutes

[1] http://research.cardiff.ac.uk/converis/portal/Project/2525952

apart, which is the ideal situation in a repeatability analysis, as hyperemia can vary through time. Unfortunately, the dataset does not contain images with and without lenses from the same checkup. It is well known that it is common for hyperemia to appear due to contact lenses use, as depicted in [10], [11]. However, in our dataset the data of the study supports that the variation in hyperemia levels between checkups one and two was too little to be significant. The reasons for this discrepancy are most likely that the previous works [10], [11] refer to a continuous exposure to contacts, such as 8-16 hours of wearing lenses, while the contact lenses comfort study established only a minimum of four hours of wear. Besides, a requirement was that all the participants must be healthy. For this reason, this dataset is suitable for the repeatability analysis.

B. Automatic hyperemia grading

Our automatic methodology comprises the steps depicted in Fig. 4. The system receives an image as input, and defines the region of interest. Then, several image features are computed and, finally, these features are combined in order to obtain the grade of the image in the given grading scale.

For the segmentation of the region of interest, split and merge segmentation [12] was applied. This method is based on a quadtree partition of the image, a data structure that consists in a tree where each parent node has exactly four children, and it is usually employed to divide in quadrants a two-dimensional space.

As a previous step, the image is thresholded with a value t. The segmentation procedure starts at the root of the tree (the initial image), evaluates a criteria of homogeneity h and, if the image does not fulfill it, this is, if the value of the criteria is higher than a threshold t_h, it is divided in four quadrants. The procedure is then repeated for each quadrant. If a quadrant fulfills the criteria, or when the minimum area a is reached, the process stops, and that quadrant becomes a leaf. If all the four quadrants from the same node are homogeneous, they are merged. The algorithm finishes when there are no more splits or merges possible. Fig. 5 depicts the obtained result.

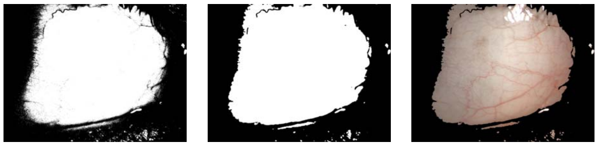

Fig. 5. Previous thresholding, mask obtained by applying split and merge segmentation, and superposition of the mask with the original image.

Then, the image features within the region of interest are computed. The features that are analyzed in this work were selected by taking into account the suggestions of the optometrists and by studying previous works [8]. The complete list is depicted in Table I. These features can be divided in two main groups, vessel-related features and hue-related features. There are 4 features belonging to the first group, and they are identified by a capital letter with the subscript v. The remaining 21 features are hue related, and they compute the intensity of a certain color in the whole conjunctiva (labeled with the capital letter I and a numeric subscript), only in the vessel area (labeled with a capital V and a numeric subscript), or only in the background of the conjunctiva (labeled with the capital letter B and a numeric subscript). The background of the conjunctiva is defined as the part of the region of interest that do not belong to any vessel.

In the formulas, the whole conjunctiva, vessel area, and background area are defined as I, VE, and $\overline{VE}$ respectively. n and m indicate the dimensions of the input image, but restricted to the pixels of the region of interest. i and j indicate the position of a given pixel (row, column). The following letters refer to the intensity value in a given channel of a certain colorspace: R, G, and B for the RGB colorspace; H, S, and V for the HSV colorspace; and L, a, and b for the L*a*b* colorspace. The feature C_v counts the number of vessels in the image, but taking into account only the values of a number of stripes n_r in the manner defined by the mask M. V_6 computes the red hue value in a similar manner as V_5, but also taking into account the neighboring pixels in a given window of size s as it is defined in the equation μ. Finally, W_v measures the average width of the vessels. To that end, a set of κ circumferences are defined, with radius ρ ranging from $n/2$ to $n/2 * \kappa$. W represents the width values for the cut points, computed using an active contour algorithm [13].

Finally, the image features are transformed to a value in the Efron scale. To that end, machine learning techniques are applied. Through previous studies, it was determined that the best results for bulbar hyperemia evaluation were obtained by using the multi-layer perceptron (MLP) [14], random forests (RF) [15], and partial least square regression (PLS) [16]. These three methods provided a better approach that other well-known state-of-art algorithms, such as support vector machines or radial-basis function networks.

C. Methodology for the study of the repeatability

Each step of the methodology was validated separately in order to remove additional bias in the results.

For the validation of the conjunctiva segmentation, a manual segmentation of the image subsets S_{clean} and S_{cont} is performed. Then, the automatic method is applied to each image, and the results are compared by counting how many pixels are labeled in the same class in both methods. A true positive is added for each pixel that both methods label as conjunctiva, a true negative when both methods label a pixel as background (non-conjunctival region), a false positive for each pixel that the automatic approach misclassifies as conjunctiva, and, finally, a false negative if the automatic approach misclassifies a pixel as background. Once these four values are obtained, the sensitivity, specificity, accuracy and precision for each image are computed, as well as the average of the whole image set. The percentage of false negatives and false positives are also calculated. These parameters give us an idea on the goodness of the method. However, this work is focused on comparing the results obtained on the *reference* image set and each *altered* image set. To that end, these

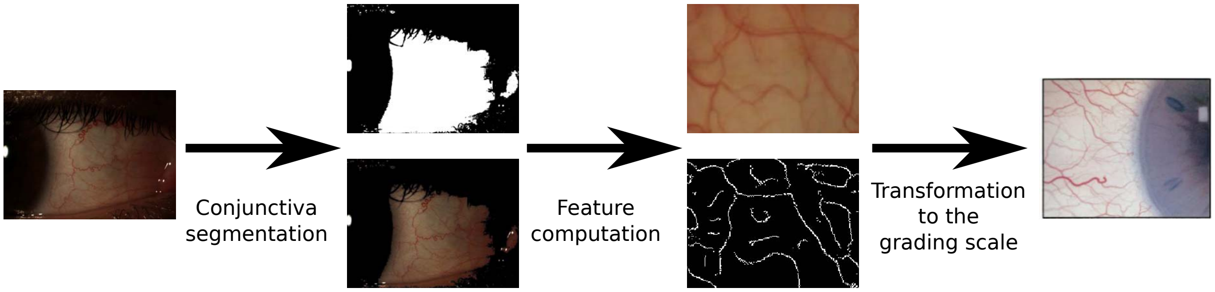

Fig. 4. Automatic methodology for bulbar hyperemia grading.

TABLE I
IMPLEMENTED HYPEREMIA FEATURES.

$\boldsymbol{B_1} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((R_{ij}+G_{ij})\overline{VE}_{ij})}{nm}$	$\boldsymbol{B_8} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((V_{ij}+S_{ij})\overline{VE}_{ij})}{nm}$	$\boldsymbol{V_6} = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{H_{ij}VE_{ij}}{\mu_{ij}}$	$\boldsymbol{I_5} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}a_{ij}}{nm}$
$\boldsymbol{B_2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((\lvert 240-H_{ij}\rvert)\overline{VE}_{ij})}{nm}$	$\boldsymbol{B_9} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((L_{ij})\overline{VE}_{ij})}{nm}$	$\mu_{ij} = \frac{\sum_{k=-s/2}^{s/2}\sum_{l=-s/2}^{s/2}\overline{VE}_{ij}Hi+k,j+l}{s^2}$	$\boldsymbol{C_v} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}E_{ij}M_{ij}}{n_r}$
$\boldsymbol{B_3} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(b_{ij}\overline{VE}_{ij})}{nm}$	$\boldsymbol{V_1} = \sum_{i=1}^{n}\sum_{j=1}^{m}(\frac{R_{ij}VE_{ij}}{R_{ij}+G_{ij}+B_{ij}})$	$\boldsymbol{V_7} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(a_{ij}VE_{ij})}{nm}$	$M_{ij} = \begin{cases} 0 & i \bmod step \neq 0 \\ 1 & i \bmod step = 0 \end{cases}$
$\boldsymbol{B_4} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(R_{ij}\overline{VE}_{ij})}{nm}$	$\boldsymbol{V_2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((R_{ij}-G_{ij})VE_{ij})}{nm}$	$\boldsymbol{I_1} = \sum_{i=1}^{n}\sum_{j=1}^{m}(\frac{R_{ij}}{R_{ij}+G_{ij}+B_{ij}})$	$\boldsymbol{A_v} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}VE_{ij}}{nm}$
$\boldsymbol{B_5} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(\lvert 128-H_{ij}\rvert)\overline{VE}_{ij})}{nm}$	$\boldsymbol{V_3} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((R_{ij}-B_{ij})VE_{ij})}{nm}$	$\boldsymbol{I_2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(R_{ij}-G_{ij})}{nm}$	$\boldsymbol{P_v} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}VE_{ij}}{nm}100$
$\boldsymbol{B_6} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((a_{ij})\overline{VE}_{ij})}{nm}$	$\boldsymbol{V_4} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}R_{ij}VE_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}VE_{ij}}100$	$\boldsymbol{I_3} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(R_{ij}-B_{ij})}{nm}$	$\boldsymbol{W_v} = \frac{\sum_{r=1}^{\rho}\sum_{c=1}^{\kappa}W_{rc}}{\rho\kappa}$
$\boldsymbol{B_7} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((R_{ij}+G_{ij}+B_{ij})\overline{VE}_{ij})}{nm}$	$\boldsymbol{V_5} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}H_{ij}VE_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}VE_{ij}}100$	$\boldsymbol{I_4} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\lvert 128-H_{ij}\rvert}{nm}$	

statistical measurements will be computed for both cases, and then compared. Ideally, the conjunctiva should be segmented with a similar success in both *reference* and *altered* cases.

Regarding the feature computation, the values of the 25 features in both images of each pair were obtained. Then, the mean and standard deviation of each feature were calculated in both sets (S_{cont} and S_{clean}), distinguishing between *reference* and *altered* images. These values were used to compute the coefficient of variation (CV), a measure commonly used in repeatability studies that provides insight on the extent of the variability relative to the mean of a population.

As the main focus of this work is to analyze the differences between the output for the *reference* and *altered* images, instead of the evaluation of the system performance, this step was started with already trained regression systems. The systems were trained with 860 images of the dataset, which were labeled by an optometrist, using 10-fold cross-validation with 100 repetitions. The 40 images that belong to S_{clean} and S_{cont} subsets were excluded, as they were used as test set. Each system received each image subset (S_{clean_ref}, S_{clean_alter}, S_{cont_ref}, and S_{cont_alter}). Then, the outputs of the two test cases S ($S = S_{clean}$ or S_{cont}) were compared with the following equation:

$$diff_S = avg((output_{S_{ref}} - output_{S_{alter}})^2) \qquad (1)$$

This is, the mean squared error (MSE) is computed by assuming that the output of S_{ref} is the expected value and the output of S_{alter} is the obtained value.

In previous works [17] the image features were studied in order to determine the best subset by means of feature selection techniques. In this article, the results obtained by the regression methods by using only the selected subsets were also analyzed. To that end, four techniques were used: two filter methods, correlation based feature selection (CFS) and Relief [18], and two wrapper methods, one using M5 algorithm (M5) [19] and another using sequential minimal optimisation for regression (SMOReg) [20].

III. RESULTS

A. Dataset

As a previous step and in order to establish our gold standard for the experiments, the variability of the optometrist evaluation in the 4 checkups for the 35 patients was studied. The expert evaluation was performed with integer and half integer values and using the Efron scale. Checkups 1 and 3 depict the eye without contact lenses, and checkups 2 to 4 depict the eye with contact lenses. The average between the nasal and tarsal shots of the eye (right and left sides) is used as ground truth. The checkups 1-2 and 3-4 are compared, as there were taken sequentially in time. For checkups 1-2, the average variation for the right eye is 0.19286, and the average variation for the left eye is 0.20714. For checkups 3-4, the average variation for the right eye is 0.19286, and for the left eye is 0.15714. 50% of the images do not vary their evaluation.

Our automatic system do not have access to the information of both sides of the eye, as it receives an image without additional knowledge regarding which side or eye it is from. Therefore, the variations of the grading in each side of each eye were also analyzed separately in order to provide a more accurate comparison. The average variation (avg), the standard deviation (std) and the percentage of images affected (% img) are depicted in Table II.

TABLE II
VARIATION OF THE EXPERTS GRADING IN THE SAME PATIENT DURING DIFFERENT CHECKUPS.

Test	Checkups(1, 2)			Checkups(3, 4)		
	Avg	Std	% img affected	Avg	Std	% img affected
RET	0.1286	0.2803	20.00	0.1571	0.2649	28.57
REN	0.2571	0.3061	45.71	0.2286	0.3286	37.14
LET	0.1714	0.2956	28.57	0.1714	0.2956	28.57
LEN	0.2429	0.3509	40.00	0.1714	0.2408	34.29

In view of the results, a certain variability in the final outputs of the system is expected between the *reference* and *altered* subsets, as even specialists have a certain level of variation. However, as the evaluations were performed using only integer and half integer values, it can be concluded that the average variation is low, as it is around 0.25 in the worst case scenario. Also, more than half of the images remain unaffected.

B. Segmentation of the conjunctiva

The values of the parameters for the automatic segmentation algorithms are $t = 40$, $h = std(I_{q_x})$ (where I_{q_x} is the intensity of the thresholded image in the quadrant q_x), $t_h = 6$, and $a = 25$. Table III depicts the average values obtained by the automatic procedure in each of the test cases. It can be observed how the parameters are similar, specially in the case of the cleaning liquid remains. However, the presence of contact lenses seems to have a bigger influence, specially on the sensitivity of the method.

TABLE III
VALIDATION OF THE REPEATABILITY OF THE ROI EXTRACTION PROCEDURE.

Set	Sens.	Spec.	Accu.	Prec.	% FN	% FP
S_{clean_ref}	0.875	0.835	0.836	0.853	0.079	0.085
S_{clean_alter}	0.851	0.833	0.819	0.839	0.091	0.090
S_{cont_ref}	0.833	0.883	0.834	0.905	0.113	0.053
S_{cont_alter}	0.805	0.905	0.820	0.921	0.138	0.042

C. Feature computation

Regarding the image features variability, Table IV depicts the obtained values for the coefficient of variation in each experiment and, also, the difference between the *reference* and *altered* sets for each test. The difference in a given set S is defined as:

$$diff_{subset} = \frac{S_{subset\_alter} - S_{subset\_ref}}{S_{subset\_alter}} \qquad (2)$$

where S_{subset} is S_{clean} or S_{cont}.

The variability of some features remain stable through all the different experiments (such as feature I_1). However, most of them react in a different manner to the different image issues. Thus, for the *cleaning* set the features that present the smallest differences between *reference* and *altered* subsets are I_1, V_6, and V_1 (ordered from lower to higher). For the *contacts* set, the smallest differences between subsets appear with features W_v, I_3, I_1, B_9, B_7, V_3, B_1, V_4, I_2, and B_3. In general, it can be observed how differences are lower in the *contacts* set. This was expected, as most of the implemented features are color-based and, therefore, they will be more affected by a change in the hue of the image, even if subtle, than by the presence of contact lenses, that is mostly irrelevant. In fact, the highest differences in the S_{cont} set take place on features A_v and P_v (both vessel quantity related measures), while the differences for the same features in S_{clean} set are much lower.

Regarding the variation of each feature in each experiment, the features can be divided in three big groups as depicted in Table V. Some features remain in the same range through all the experiments, such as V_4, B_1, B_4, B_7, B_9, C_v, V_1, and V_6. The range of these features is less likely to be affected by image conditions such as the presented. Therefore, the features that are related to the hue in the background are less affected than the ones that take into account the vessels, specially in the subset S_{clean}.

D. Transformation to the scale value

By looking at the results provided by the feature selection techniques, depicted in Table VI, it can be observed that several of the selected features vary their range depending on the experiment. Therefore, to obtain appreciable differences between the *reference* and *altered* subsets is expected. Note that only one vessel-related feature was taken into account, W_v in the SMOReg approach.

TABLE VI
FEATURES THAT APPEAR IN AT LEAST 7 OUT OF 10 FOLDS.

Method	#	selected features
CFS	12	V_1, V_2, I_2, V_3, I_4, I_5, V_7, B_2, B_5, B_6, B_7, B_9
Relief	8	I_1, I_3, I_4, V_6, V_7, B_2, B_3, B_5
M5	7	V_1, V_3, I_3, I_4, V_5, V_7, B_9
SMOReg	13	V_1, V_2, V_3, I_3, I_4, V_5, V_6, V_7, B_1, B_2, B_5, B_9, W_v

TABLE IV
COEFFICIENT OF VARIATION FOR EACH FEATURE COMPARED TO THE EXPERT.

Feature	CV S_{clean_ref}	CV S_{clean_alter}	diff(S_{clean})	CV S_{cont_ref}	CV S_{cont_alter}	diff(S_{cont})
B_1	0.0600	0.1121	0.8684	0.1081	0.1181	0.0928
B_2	0.0242	0.2736	10.297	0.1017	0.0344	0.6621
B_3	0.3194	1.1443	2.5830	0.3514	0.3145	0.1050
B_4	0.0537	0.1054	0.9647	0.1029	0.1153	0.1213
B_5	0.0260	0.2433	8.3400	0.0244	0.0095	0.6101
B_6	0.2661	1.2209	3.5879	0.2459	0.3057	0.2431
B_7	0.0694	0.1254	0.8073	0.1152	0.1240	0.0763
B_8	0.1767	0.2121	0.2002	0.2353	0.2657	0.1291
B_9	0.0622	0.1111	0.7881	0.1069	0.1145	0.0717
V_1	0.5210	0.5765	0.1065	0.4333	0.5871	0.3550
V_2	0.1697	0.4422	1.6063	0.2010	0.2468	0.2283
V_3	0.2096	0.5385	1.5695	0.2460	0.2649	0.0770
V_4	0.0611	0.0981	0.6051	0.1112	0.1219	0.0963
V_5	0.1841	0.8245	3.4778	1.3482	0.5057	0.6249
V_6	1.2612	1.1858	0.0597	1.2636	1.0633	0.1585
V_7	0.1633	0.4354	1.6660	0.1885	0.2516	0.3349
I_1	0.3054	0.3002	0.0169	0.2383	0.2255	0.0536
I_2	0.2531	1.0782	3.2593	0.2601	0.2857	0.0983
I_3	0.2749	1.0619	2.8632	0.3004	0.2921	0.0278
I_4	0.0260	0.2392	8.2012	0.0241	0.0094	0.6083
I_5	0.2645	1.1895	3.4973	0.2461	0.3057	0.2423
C_v	0.4899	0.6530	0.3328	0.4763	0.5644	0.1851
A_v	0.3481	0.4878	0.4013	0.2501	0.4388	0.7544
P_v	0.3481	0.4878	0.4013	0.2501	0.4388	0.7544
W_v	0.0823	0.3683	3.4752	0.0810	0.0799	0.0133

TABLE V
FEATURES GROUPED BY COEFFICIENT OF VARIATION.

% of variation	S_{clean_ref}	S_{clean_alter}	S_{cont_ref}	S_{cont_alter}
$\leq 20\%$	$V_2, I_4, V_4, V_5, V_7, B_1$ $B_2, B_4, B_5, B_7, B_8, B_9$	V_4, B_1, B_4, B_7, B_9	I_4, V_4, V_7, B_1, B_2 B_4, B_5, B_7, B_9	I_4, V_4, B_1, B_2 B_4, B_5, B_7, B_9
$20\% - 30\%$	I_2, V_3, I_3, I_5, B_6	I_4, B_2, B_5, B_8	A_v, I_1, V_2, I_2, V_3 P_v, I_5, B_6, B_8	I_1, V_2, I_2, V_3 I_3, V_7, B_6, B_8
$30\% - 40\%$	A_v, I_1, P_v, B_3		I_3, B_3	I_5, B_3
$\geq 40\%$	C_v, V_1, V_6	C_v, A_v, V_1, I_1, V_2 I_2, V_3, I_3, P_v, V_5 V_6, I_5, V_7, B_3, B_6	C_v, V_1, V_5, V_6	$C_v, A_v, V_1, P_v, V_5, V_6$

TABLE VII
PARAMETERS FOR THE REGRESSION TECHNIQUES.

Method	Parameters
MLP	configuration = [40 16] activation function = hyperbolic tangent sigmoid training function = Bayesian regularization backpropagation based on Levenberg-Marquardt optimization epochs = 1000 weight initialization = Nguyen-Widrow
PLS	number of components = min(number of features, 8)
RF	number of trees = 60 minimum leaf size = 10

The parameters for the regression methods were chosen empirically, and they are depicted in Table VII. The MSE results for the analyzed machine learning techniques and each test set are depicted in Table VIII. It can be observed how the differences in the results for the RF approach are minimal, even in the S_{clean} set. Both MLP and PLS seem to be the most affected by the blue hue of the cleaning liquid test, as the values for S_{clean} set are much worse than the ones for S_{cont} set in all the cases, specially with PLS when using all the features.

Additionally, the MSE values for the whole image set, obtained by averaging the test error in the k-fold, are shown in Table IX. The systems that obtain the lowest MSE are the PLS and the RF with CFS and SMOReg feature subsets, respectively. By comparing these results with the ones in Table VIII, it can be observed how these approaches provide usually low differentiation in both S_{clean} and S_{cont} and, therefore, the systems provide an evaluation similar to the optometrist's.

Other goodness metrics were also computed, such as the mean absolute error (MAE) or the coefficient of determination (R^2). The systems that obtained the best values were the same, although the order vary depending on the metric. PLS with CFS obtained a $MAE = 0.2719$ and $R^2 = 0.3278$, while RF with SMOReg obtained a $MAE = 0.2756$ and $R^2 = 0.5899$.

In order to better observe the results for each regression technique with its best feature set, Fig. 6 depicts the scatter plots for predicted and real values in the three cases. Moreover, a statistical test was conducted in order to assess if the

TABLE VIII
MSE FOR EACH COMBINATION OF FEATURES SET AND REGRESSION TECHNIQUE.

All

Set	MLP	PLS	RF
S_{clean}	0.3551	0.1362	0.0666
S_{cont}	0.3255	0.0309	0.0275

CFS

Set	MLP	PLS	RF
S_{clean}	0.4999	0.0972	0.0708
S_{cont}	0.3764	0.0609	0.0396

Relief

Set	MLP	PLS	RF
S_{clean}	0.1424	0.1055	0.0681
S_{cont}	0.0813	0.0827	0.0723

M5

Set	MLP	PLS	RF
S_{clean}	0.3658	0.0838	0.0756
S_{cont}	0.2477	0.0455	0.0380

SMOReg

Set	MLP	PLS	RF
S_{clean}	0.1731	0.0945	0.0685
S_{cont}	0.1093	0.0454	0.0365

TABLE IX
MSE FOR THE WHOLE DATASET AND EACH (FEATURE SELECTION, REGRESSION TECHNIQUE) COMBINATION.

MSE

Features	MLP	PLS	RF
All	0.2826	0.1313	0.1194
CFS	0.4040	0.1166	0.1184
Relief	0.1871	0.1323	0.1262
M5	0.3095	0.1176	0.1182
SMOReg	0.2340	0.1196	0.1175

differences between the automatic outputs for the best system and the expert's evaluations are significant. A normality test was performed in both samples, obtaining a strong reject of the null hypothesis. Therefore, a Wilcoxon signed rank test for the differences was conducted, and the null hypothesis (the difference comes from a distribution with zero median) was accepted with $\alpha = 0.05$ and a p-value of 0.1153.

In order to ensure that our systems perform in a similar manner than a human expert, the variability of the optometrist's evaluation between the checkups was studied and compared to the evaluations provided by the automatic approaches. The clinical study reported that these differences between the checkups' evaluations were low, but it must be confirmed that the results from our systems are within the same range. Therefore, the outputs of our regression systems for the 900 images of the dataset with the full feature set were obtained, and the average differences between checkups were computed. The differences were computed pairwise, using the checkups that are consecutive. Therefore, for a certain combination of patient, eye, and side, the difference on its evaluation between checkups 1-2, between checkups 2-3, and between checkups 3-4 was computed. This value was also computed for the manual evaluations in order to establish a comparison. Table X depicts the obtained results for both the average difference and the standard deviation. The results are not expressed with the coefficient of variation as the compared values are in the same range, so both mean and standard deviation can be directly compared. Besides, the means are too close to zero in some of the tests, and the individual observations have different sign, which can cause the parameter to be misleading.

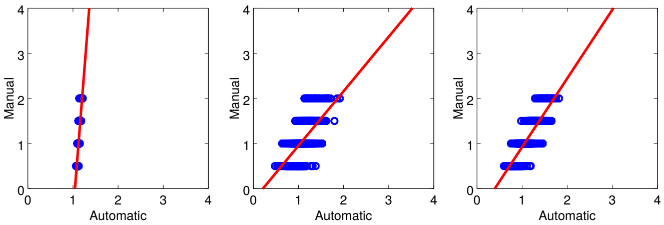

Fig. 6. Scatter plots for each system with their best subset. Left to right: MLP with Relief, PLS with CFS and RF with SMOReg.

TABLE X
DIFFERENCES ON THE EVALUATION OF THE SAME CASE THROUGH DIFFERENT CHECKUPS.

System	MLP	PLS	RF	Manual
Avg. diff(C1,C2)	0.0198	0.0061	0.0318	0.0772
Avg. diff(C2,C3)	-0.0255	-0.0188	-0.0623	-0.1047
Avg. diff(C3,C4)	0.0068	0.0199	0.0414	0.0367
Std. diff(C1,C2)	0.1405	0.2339	0.2980	0.4670
Std. diff(C2,C3)	0.1301	0.2327	0.2791	0.5578
Std. diff(C3,C4)	0.1211	0.2378	0.2757	0.5276

It can be observed how the systems' values are similar to the manual evaluations. All the automatic approaches have a lower standard deviation, and some of the systems present a lower mean value for a certain pair of checkups. Additionally, those cases that the expert graded with the same values in a pair of checkups were analyzed, and the magnitude of the system variation was observed (Table XI). 96, 92 and 92 images present no variations in a given pair of checkups, 1-2, 2-3, and 3-4 respectively. The results were good, as the average differences are closer to zero. The best values, taking into account the combination of mean and standard deviation, are achieved by the MLP in all pairs. The overall worst results are obtained by PLS.

TABLE XI
MAGNITUDE OF THE VARIATION IN THE AUTOMATIC SYSTEMS FOR THOSE CASES WHERE THE MANUAL EVALUATION DOES NOT VARY.

System	MLP	PLS	RF
Avg. diff(C1,C2)	0.0013	-0.0135	-0.0140
Avg. diff(C2,C3)	0.0043	0.0558	0.0330
Avg. diff(C3,C4)	-0.0063	-0.0034	-0.0013
Std. diff(C1,C2)	0.0209	0.1611	0.1508
Std. diff(C2,C3)	0.0175	0.2115	0.1505
Std. diff(C3,C4)	0.0172	0.2049	0.1401

Finally, a study was performed in order to know if the sign of the systems' variations was the same that in the expert's one in those cases where the expert's evaluation varied (Table XII). The best results are obtained by the MLP, that achieves the same sign in its variation than the human expert in all the cases. The PLS is much more inconsistent, while the RF obtains also good results.

TABLE XII
ANALYSIS OF THE SIGN OF THE VARIATION IN MANUAL AND AUTOMATIC CASES, FOR THE VALUES WHERE THE MANUAL EVALUATION VARIES.

System	MLP	PLS	RF
% same sign (C1,C2)	100.00	73.30	91.48
% same sign (C2,C3)	100.00	71.43	96.27
% same sign (C3,C4)	100.00	76.19	95.24

IV. CONCLUSIONS

In this work, the effect that different image alterations have in our automatic hyperemia grading methodology was analyzed. As the inputs of the system have a high variability, there is a need to ensure that the system's performance is only affected in the same manner as a human expert. Therefore, each step of the methodology was tested separately under two different alterations: the presence of contact lenses or the remains of a blue cleaning lotion.

The first step, the segmentation of the region of interest, is more affected by the presence of contact lenses. However, the variation was low in all tests. The feature computation is more affected by the changes in hue, as most of the features are color based. Therefore, the largest differences were obtained when testing the images with remains of blue lotion. The features computing the background intensity are more stable in general. Regarding the regression techniques used to combine the image features into the final grade in the scale, the most stable approach is the RF. However, further analysis on specific cases shows that the MLP tends to not change its evaluation when the optometrist also does not change and, when the evaluation changes, the MLP almost always maintains the same sign of the optometrist's change. Finally, the systems' differences between consecutive checkups and the expert's one were compared, and it can be concluded that the values are similar and, therefore, that the variations in the inputs of the system produce a variation in the result that is within the same range as the optometrist.

Our future lines of work include the integration of the methodology as a part of an assisted diagnosis tool, and the study of the evolution of a patient through different evaluations in order to track the progression of the symptom.

ACKNOWLEDGMENT

This research has been partially by the Ministerio de Economía y Competitividad through the research contract DPI2015-69948-R. María Luisa Sánchez Brea acknowledges the support of the University of A Coruna though the Inditex-UDC Grant Program.

REFERENCES

[1] H. Cronau, R. R. Kankanala, and T. Mauger, "Diagnosis and management of red eye in primary care," *Am Fam Physician*, vol. 81, no. 2, pp. 137–144, 2010.

[2] M. Rolando and M. Zierhut, "The ocular surface and tear film and their dysfunction in dry eye disease," *Survey of ophthalmology*, vol. 45, pp. S203–S210, 2001.

[3] M. L. S. Brea, N. B. Rodríguez, N. S. Maroño, A. M. González, C. García-Resúa, and M. J. G. Fernández, "On the development of conjunctival hyperemia computer-assisted diagnosis tools: Influence of feature selection and class imbalance in automatic gradings," *Artificial Intelligence in Medicine*, vol. 71, pp. 30–42, 2016.

[4] J. D. Rodriguez, P. R. Johnston, G. W. Ousler III, L. M. Smith, and M. B. Abelson, "Automated grading system for evaluation of ocular redness associated with dry eye," *Clinical ophthalmology (Auckland, NZ)*, vol. 7, p. 1197, 2013.

[5] 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, 2012.

[6] J. S. Wolffsohn and C. Purslow, "Clinical monitoring of ocular physiology using digital image analysis," *Contact Lens and Anterior Eye*, vol. 26, no. 1, pp. 27–35, 2003.

[7] 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, 2013.

[8] E. B. Papas, "Key factors in the subjective and objective assessment of conjunctival erythema," *Investigative ophthalmology & visual science*, vol. 41, no. 3, pp. 687–691, 2000.

[9] L. Sánchez, N. Barreira, H. Pena-Verdeal, and E. Yebra-Pimentel, "A novel framework for hyperemia grading based on artificial neural networks," in *International Work-Conference on Artificial Neural Networks*. Springer, 2015, pp. 263–275.

[10] J. Walker, G. Young, C. Hunt, and T. Henderson, "Multi-centre evaluation of two daily disposable contact lenses," *Contact Lens and Anterior Eye*, vol. 30, no. 2, pp. 125–133, 2007.

[11] J. S. Wolffsohn, O. A. Hunt, and A. Chowdhury, "Objective clinical performance of comfort-enhanceddaily disposable soft contact lenses," *Contact Lens and Anterior Eye*, vol. 33, no. 2, pp. 88–92, 2010.

[12] H.-D. Cheng, X. Jiang, Y. Sun, and J. Wang, "Color image segmentation: advances and prospects," *Pattern recognition*, vol. 34, no. 12, pp. 2259–2281, 2001.

[13] S. Vázquez, N. Barreira, M. G. Penedo, M. Pena-Seijo, and F. Gómez-Ulla, "Evaluation of SIRIUS retinal vessel width measurement in REVIEW dataset," in *CBMS*, 2013, pp. 71–76.

[14] S. Mehrabi, M. Maghsoudloo, H. Arabalibeik, R. Noormand, and Y. Nozari, "Application of multilayer perceptron and radial basis function neural networks in differentiating between chronic obstructive pulmonary and congestive heart failure diseases," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6956–6959, 2009.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] H. Abdi, "Partial least square regression (PLS regression)," *Encyclopedia for research methods for the social sciences*, pp. 792–795, 2003.

[17] L. Sánchez, N. Barreira, N. Sánchez-Maroño, A. Mosquera, C. García-Resúa, and E. Yebra-Pimentel, "On the analysis of feature selection techniques in a conjunctival hyperemia grading framework," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2016, pp. 271–276.

[18] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[19] D. P. Solomatine and Y. Xue, "M5 model trees and neural networks: application to flood forecasting in the upper reach of the huai river in china," *Journal of Hydrologic Engineering*, vol. 9, no. 6, pp. 491–501, 2004.

[20] S. Shevade, S. Keerthi, C. Bhattacharyya, and K. Murthy, "Improvements to the smo algorithm for svm regression," *Neural Networks, IEEE Trans. on*, vol. 11, no. 5, pp. 1188–1193, 2000.