The Use of Datasets of Bad Quality Images to Define Fundus Image Quality

Matteo Menolotto, *Member, IEEE,* Mario E. Giardini, *Member, EMBS*

***Abstract*— Screening programs for sight-threatening diseases rely on the grading of a large number of digital retinal images. As automatic image grading technology evolves, there emerges a need to provide a rigorous definition of image quality with reference to the grading task. In this work, on two subsets of the CORD database of clinically gradable and matching non-gradable digital retinal images, a feature set based on statistical and on task-specific morphological features has been identified. A machine learning technique has then been demonstrated to classify the images as per their clinical gradeability, offering a proxy for a rigorous definition of image quality.**

***Clinical Relevance*— This work offers a novel strategy to define fundus image quality, to contribute to the development of automatic fundus image graders for retinal screening.**

I. Introduction

Reducing the economic and social impact of avoidable blindness and vision impairment, particularly severe in low- and middle- income countries, has been identified as a key action by the World Health Organization [1]. Almost half billion people suffer from treatable sight pathologies that, however, show symptoms only in their late stages, such as diabetic retinopathy, age related macular degeneration and glaucoma [1]. The most effective prevention tools are population-wide screening programs, which, however, produce a large quantity of retinal images that need to be graded for pathological markers, creating a bottleneck in the availability of professional staff trained to do so. Automatic software able to distinguish between healthy and non-healthy retinas are starting to be employed by the public health service [2] to unburden some stages of such grading process.

The quality of the digital retinal images has a major impact on the classification performance of automatic screening tools. Yet, for this task, a formal definition of image quality is still elusive. In clinical practice, ophthalmologists rely on their experience and knowledge to determine whether the clinical content of an image is adequate to formulate a diagnosis. Such decision-making process involves several complex cognitive tasks [3], which makes it very difficult to relate this quality definition strategy to an image processing tool. Indeed, the definition of objective quality in fundoscopic images is still a matter under very active debate [4], and yet necessary in a rigorous approach to high-throughput automatic retinal image classifiers.

To a certain extent, the clinical content of a retinal image is associated with its textural content, which in turn is related with morphology (relation to geometrical structuring elements) and chromaticity. A typical retinal image, in fact, contains many anatomical structures, such as blood vessels and optic nerve hea*d, and may include features that can be associated to pathological conditions, such as dark and bright lesions. Abdel-Hamid *et al.* [5] implemented a quality assessment algorithm that evaluates textural-related elements such as sharpness and homogeneity. In the work of Fu *et al.* [6], quality-related features where evaluated on different color spaces and combined to train a deep learning network. However, artifacts, noise and distortions can contribute to add textural elements to the image, making quality classification based on texture complicated, and possibly ill-defined. Nonetheless, the understandable desire to link a formal quality definition to the clinical information content of a retinal image is stimulating the search for other definition criteria. Dias *et al.* [7] proposed classic photography-related indicators such as color, focus, contrast, and illumination, to distinguish between gradable and ungradable retinal images. Although the sensitivity reached over 97%, the classifier was mainly trained to detect over- (bright) and under-exposed (dark) retinal images. More recent feature-specific quality descriptors quantify the amount of a specific anatomical feature in the retinal image. The majority of them are based on segmentation techniques, e.g., to quantify the amount of blood vessels [8, 9] or the visibility of the optic disc [10, 11]. However, once again, these are prone to errors caused by artifacts and distortions.

In this work, we develop a proxy of image quality, using a quality classifier ultimately based on how selected feature and detail metrics of the images are affected by artifacts. These metrics can be reasonably expected to better correlate to information content that e.g., simple brightness / darkness. To achieve this, we use the unique features provided by the open access CORD database [12], which includes images of the retina with clinical gradable quality, alongside the same images with template artifacts, to train a machine learning classifier to estimate whether images are gradable or not, based on classic image quality-related statistical indicators, and simple retinal image-specific quantifiers.

II. Retinal Image Parametrization

Common quality-related parameters used in photography are statistical descriptors based on histogram and Haralick features [13]. In this work, eleven of such parameters, along with other two parameters that highlight specific anatomical

Research supported by the Rosetrees Trust, UK (grant M720) and Sight Research UK, former National Eye Research Centre (grant SAC030).

M. Menolotto is with the Tyndall National Institute, University College Cork, T12 R5CP Cork, Ireland (corresponding author e-mail: matteo.menolotto@tyndall.ie)

M. E. Giardini is with the University of Strathclyde, G1 1XQ, Glasgow, Scotland, UK (mario.giardini@strath.ac.uk)

retinal features, are evaluated on the two subsets of retinal images included in CORD.

A. Histogram and Contrast Features

The six histogram features selected for our study are: mean, standard deviation, skewness, kurtosis, interquartile range (IQR) and contrast sensitivity function (CSF), where the CSF of a channel X is obtained as:

$$CSF(X) = IQR(X)/\,max(X) - min(X), \tag{1}$$

and is an expression of the statistical dispersion of the between the upper and lower quartile respect to the range of intensities of that channel [14].

To account for uneven illumination and poor focus, seven different contrast and blur parameters have been selected: contrast ratio (CR), local contrast ratio (LCR), blur metric, full intensity range (R), relative intensity range and saturation metrics [15, 16]. Contrast ratio is calculated as:

$$CR_j = \bar{p}_j / s_j, \tag{2}$$

where $\bar{p}_j$ is the mean intensity of all of the pixels in a region of interest (ROI), in the channel j while s_j is the standard deviation of the pixels in the same ROI in the channel j. The ROI, in this case, is the whole retinal image excluding the black borders (Fig. 1). The higher the blurriness, the higher the CR. A similar contrast indicator is the local contrast ratio, which is the CR calculated on non-overlapping sub-windows of the retinal image as follows:

$$LCR = \left(\sum_{i=1}^{n} \frac{\bar{p}_{w,i}}{s_{w,i}}\right)/n, \tag{3}$$

where w is a $N \times N$ window inside the ROI and n is the total number of sub-windows.

The blur metric measures the focal blur and the motion blur by comparing the original image with its low-pass filtered version. The intensity range measures the grayscale spread of the image. A larger range usually indicates higher contrast in an image. As saturation metrics, the proportion of pixels at the highest (P_{max}) and lowest (P_{min}) intensity level are computed, which can reveal over- or underexposure respectively.

B. Haralick Features

Texture, along with spectrum and context are the three fundamental pattern elements used in human interpretation of color images. Haralick at al. developed a classification system for texture based on the statistical evaluation not of the image itself but, rather, of grey-tone spatial-dependence matrices obtained from it [13]. This method is based on the assumption that grey tone and texture have a mutual interconnection to one another, thus highlighting the complexity of the grey tone transitions within the image, revealing the presence of organized structures or homogeneity, and the prevalence between texture and tone. Haralick et al. identified 14 different textural features of which, in this work, we considered 5:

- Energy: $H_1 = \sum_i \sum_j \{p(i,j)\}^2$
- Contrast: $H_2 = \sum_{n=0}^{N-1} n^2 \left\{\sum_{i=1}^{N} \sum_{j=1}^{N} p(i,j) \,\middle|\, |i-j| = n\right\}$
- Correlation: $H_3 = \frac{\sum_i \sum_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
- Homogeneity: $H_4 = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$
- Entropy: $H_5 = -\sum_i \sum_j p(i,j) \log\big(p(i,j)\big)$

Where $p(i,j)$ is the (i,j)th entry in a normalised co-occurrence matrix P. μ_x, μ_y, σ_x and σ_y are the mean and standard deviation of p_x and p_y respectively, which represent the marginal-probability matrix obtained as $\sum_{j=1}^{N} P(i,j)$, where the number of distinctive grey level is $n = 1, \ldots, N$.

B. Retinal-specific Textural Features

In this work, two parameters specifically linked to retinal images were included, namely blood vessel density (BVD) and blood vessel contrast (BVC). To compute these two values for each retinal image, the blood vessels were firstly isolated, generating a binary map M where blood vessel and the background have two different value (0 and 1), using a vessel segmentation technique based on a matching filter algorithm [8], implementing a kernel with 12 different orientations (rotation of 15° steps) and fixing an arbitrary threshold T, as follows:

$$M(i,j) = \begin{cases} 1, & \tilde{g}(i,j) > T \\ 0, & otherwise \end{cases} \tag{4}$$

where $\tilde{g}$ is the result of the filtering on the histogram equalized green channel g of the retinal image. BVD is the ratio between the number of pixels that belongs to the blood vessels and to the total amount of pixels in the ROI, expressed as:

$$BVD = \frac{\sum_{i=1,j=1}^{m,n} M(i,j)}{m \times n}, \tag{5}$$

where m and n are the width and height of the image in pixels respectively. Blood vessel contrast is defined as the contrast of the pixels of the blood vessels with respect to the background, and is obtained using the following:

$$BVC = |\bar{p} \in M(i,j) - \bar{p} \notin M(i,j)| \tag{6}$$

III. METHODS

The retinal image set in CORD consists of 548 fundus images acquired via fundus camera (FC), and 80 optical coherence tomography (OCT) scans, each also associated with a set of monocular and stereoscopic fundus images captured through the OCT instrument optics itself. The CORD database also contains 231 photos and 160 videos from slit lamp examination, not used in this study. Excluding the fundus images acquired with the OCT instrument in stereo imaging modality, the total amount of retinal images acquired via F, and via the OCT instrument, and divided in "clinical standard" quality (CSQ) and artifact, is summarized in Table I. Example of CSQ and artifact macula-centered retinal images are shown in Fig. 1. Twenty different statistical features have been evaluated on the RGB, HVI and CIELab color spaces of the two datasets of CORD, the CSQ and the artifact affected fundus images. The machine learning classifier trained with this data is a diagonal adaptation of Neighborhood Component Analysis (NCA) [17, 18]. This

algorithm is also able to identify each feature significance to the classification process.

TABLE I. FUNDUS IMAGES IN CORD

Fundus imaging instrument	Artifact	CSQ	Total
FC[a]	251	37	288
OCT[a]	40	20	60

a. Images subjected to complete object obstruction were excluded from the total dataset.

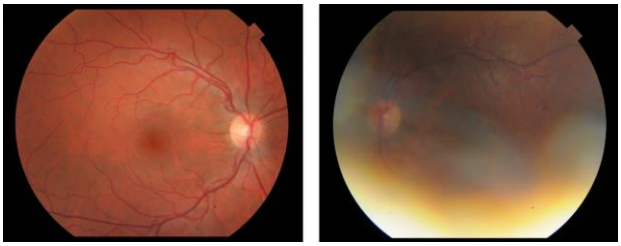

Figure 1. Retinal images acquired using a fundus camera Topcon TRC-50DX Type IA fitted with the body of a Nikon D300s. Left: clinical standard quality retinal image. Right: artifact caused by patient movement.

To establish the best channel for quality classification the datasets of images captured via FC and via the OCT instrument were split into two parts: the classification training subset and the test subset. The training process of the classifier started by using 1/8th of the total retinal images available (starting training subset) and was increased by 1/8th of the total retinal images available until the classification process was able to correctly classify the remaining images (test subset) (Fig. 2). After identifying the most sensitive channel for quality classification, features with a NCA weight ≥ 0.4 were considered for the clustering. The more the weight of the feature, the stronger the influence in the classification process.

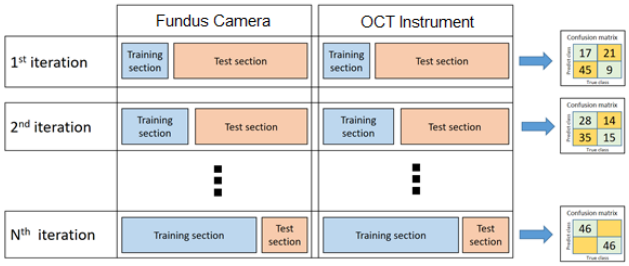

Figure 2. Visual description of the iteration process used to find the minimum training dataset able to classify the test subset correctly.

IV. RESULTS

The color channels which model shown the best prediction rate were the second channel of the CIELab color space as for the FC images, and the green channel of the RGB color space for the images acquired via the OCT instrument (Table II). For the images captured via FC, BVC and IQR resulted the best features for classification in the majority of the channels tested, with the first being relevant in all but the intensity channel of the HIS color space.

TABLE II. BEST PREDICTIVE FEATURES FOR QUALITY CLASSIFICATION

Fundus imaging instrument	Best color space (channel)	Most relevant features
FC	CIELab (a)	IQR, BVC, Mean
OCT	RGB (Green)	Kurtosis, Mean, R

On the contrary, for the images captured via the OCT instrument, range and kurtosis were the most significant overall. The mean value of the pixel intensity of the resulting best channels had also a major role in the quality classification in both imaging techniques. The cluster plots for the two imaging instruments are shown in Fig. 3 and Fig. 4.

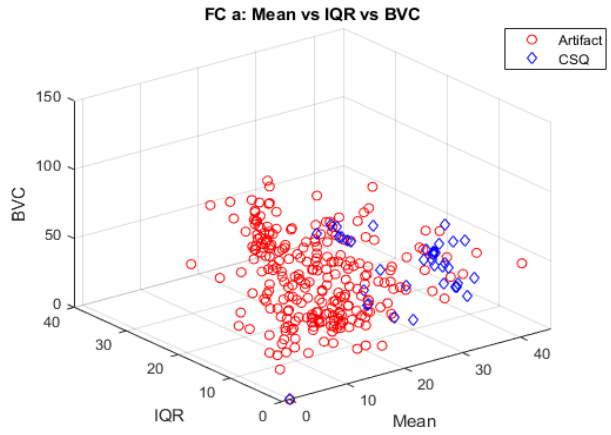

Figure 3. 3D scatter plot of the best three features used to cluster FC images.

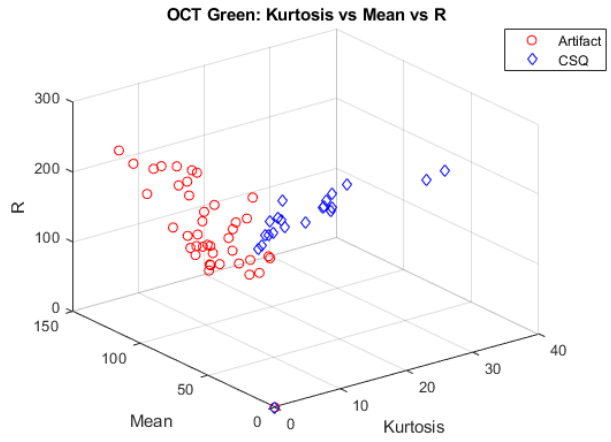

Figure 4. 3D scatter plot of the best three features used to cluster images captured via the OCT instrument.

V. DISCUSSION

The classification models selected for the images captured via the FC and the via OCT instrument show better classification using the triplet [IQR, mean, BVC] calculated for the a channel of the CIELab color space and [mean, range, kurtosis] calculated for the $green$ channel of the RGB color space, respectively. The good performance shown by specific textural features was somehow predictable, given that the anatomical features (blood vessels, optic disk and macula),

which are embedded in texture, are what mostly characterizes the information of a retinal image. Less expected was the significant contribution to classification performance of common histogram features, such as mean value and range as, in general, histograms give information about the general aspect of the whole image, rather than local information content. Histograms appear to differentiate images where local information is key, such as those with good anatomical features, from images with artifacts or photographic defects.

The decision of analyzing separately the images obtained via FC and OCT (albeit used in FC mode) was made to account for the different optics and settings of the two instruments and of the related artifact generation in CORD, which appears to be more repeatable for the artifacts generated on the OCT system than on the FC. A high level of repeatability boosts the identification of specific patterns in the images, hence improve the classification process. Therefore, the clustering of the fundus images generated via the OCT instrument is better than for the FC images (Fig. 3 VS Fig. 4). As for all machine learning classifiers, we would expect the clustering performance to improve as the amount and diversity of training data increases, advocating for the creation of more databases containing examples of artifacts. Indeed, at present our dataset contains images from 10 healthy subjects only, and increasing diversity, and including pathology, may help in improving generalizability.

VI. Conclusion

Based on an established machine learning technique, a small set of digital image features to classify retinal images as gradable vs. non-gradable have been identified for two imaging instruments. Such features include both classic photographic quality indicators, and retinal-specific features, denoting that the combination of these two types yields a more significant image quality classification, whether the retinal image is affected by quality distortions caused by camera settings (e.g. defocusing, overexposure) or by common fundoscopy artifacts. The method has been enabled by the availability of CORD, a dataset of retinal images that contain images of gradable quality, and their counterparts with artifacts. This work highlighted how creating datasets containing images with quality degradations can underpin a strategy for defining image quality in fundoscopy.

Future work will focus on two main objectives: expanding CORD with more example of artifacts and quality degradations, at the same time increasing diversity and, possibly, extending it to pathology, and the implementation of different classifiers, possibly identifying different image parametrizations, to match a more objective definition of image quality in fundoscopy, e.g., related to task-specific performance.

Finally, with this work we aimed to demonstrate the importance of dataset of “bad quality” retinal images alongside their “good quality” counterparts, as a way to better understand the impact of artifacts and common degradations in fundoscopy on the clinical content of the images.

Acknowledgment

The authors would like to thank Dr. Iain Livingstone for his support and help during the development of this research.

References

[1] W. H. Organization, "World report on vision," 9241516577, 2019.

[2] G. S. Scotland *et al.*, "Cost-effectiveness of implementing automated grading within the national screening programme for diabetic retinopathy in Scotland," *Br J Ophthalmol,* vol. 91, no. 11, pp. 1518-23, Nov 2007.

[3] D. Manning, S. Ethell, T. Donovan, and T. Crawford, "How do radiologists do it? The influence of experience and training on searching for chest nodules," *Radiography,* vol. 12, no. 2, pp. 134-142, 2006.

[4] M. Lalonde, L. Gagnon, and M.-C. Boucher, "Automatic visual quality assessment in optical fundus images," in *Proceedings of vision interface*, 2001, vol. 32, pp. 259-264: Ottawa.

[5] L. Abdel-Hamid, A. El-Rafei, S. El-Ramly, G. Michelson, and J. Hornegger, "Retinal image quality assessment based on image clarity and content," *J Biomed Opt,* vol. 21, no. 9, p. 96007, Sep 1 2016.

[6] H. Fu *et al.*, "Evaluation of Retinal Image Quality Assessment Networks in Different Color-Spaces," vol. 11764, pp. 48-56, 2019.

[7] J. M. P. Dias, C. M. Oliveira, and L. A. d. S. Cruz, "Evaluation of Retinal Image Gradability by Image Features Classification," *Procedia Technology,* vol. 5, pp. 865-875, 2012.

[8] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of blood vessels in retinal images using two-dimensional matched filters," *IEEE Trans Med Imaging,* vol. 8, no. 3, pp. 263-9, 1989.

[9] J. V. Soares, J. J. Leandro, R. M. Cesar Junior, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *IEEE Trans Med Imaging,* vol. 25, no. 9, pp. 1214-22, Sep 2006.

[10] P. S. Mittapalli and G. B. Kande, "Segmentation of optic disk and optic cup from digital fundus images for the assessment of glaucoma," *Biomedical Signal Processing and Control,* vol. 24, pp. 34-46, 2016.

[11] G. D. Joshi, J. Sivaswamy, and S. R. Krishnadas, "Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment," *IEEE Trans Med Imaging,* vol. 30, no. 6, pp. 1192-205, Jun 2011.

[12] M. Menolotto, K. Jordan, I. Coghill, and M. E. Giardini, "CORD - Comprehensive Ophthalmic Research Database," ed. https://pureportal.strath.ac.uk/en/datasets/comprehensive-ophthalmic-research-database-cord: University of Strathclyde, 2018.

[13] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. SMC-3, no. 6, pp. 610-621, 1973.

[14] K. Seshadrinathan *et al.*, "Image Quality Assessment," 2009, pp. 553-595.

[15] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: Perception and estimation with a new no-reference perceptual blur metric," (in English), *Human Vision and Electronic Imaging Xii,* vol. 6492, 2007.

[16] F. Yin *et al.*, "Automatic retinal interest evaluation system (ARIES)," *Conf Proc IEEE Eng Med Biol Soc,* vol. 2014, pp. 162-5, 2014.

[17] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in neural information processing systems*, 2005, pp. 513-520.

[18] W. Yang, K. Wang, and W. Zuo, "Neighborhood Component Feature Selection for High-Dimensional Data," *JCP,* vol. 7, no. 1, pp. 161-168, 2012.