

Human Visual System-Based Fundus Image Quality Assessment of Portable Fundus Camera Photographs

Shaoze Wang, Kai Jin, Haitong Lu, Chuming Cheng, Juan Ye*, and Dahong Qian*, *Senior Member, IEEE*

Abstract—Telemedicine and the medical “big data” era in ophthalmology highlight the use of non-mydratric ocular fundus photography, which has given rise to indispensable applications of portable fundus cameras. However, in the case of portable fundus photography, non-mydratric image quality is more vulnerable to distortions, such as uneven illumination, color distortion, blur, and low contrast. Such distortions are called generic quality distortions. This paper proposes an algorithm capable of selecting images of fair generic quality that would be especially useful to assist inexperienced individuals in collecting meaningful and interpretable data with consistency. The algorithm is based on three characteristics of the human visual system—multi-channel sensation, just noticeable blur, and the contrast sensitivity function to detect illumination and color distortion, blur, and low contrast distortion, respectively. A total of 536 retinal images, 280 from proprietary databases and 256 from public databases, were graded independently by one senior and two junior ophthalmologists, such that three partial measures of quality and generic overall quality were classified into two categories. Binary classification was implemented by the support vector machine and the decision tree, and receiver operating characteristic (ROC) curves were obtained and plotted to analyze the performance of the proposed algorithm. The experimental results revealed that the generic overall quality classification achieved a sensitivity of 87.45% at a specificity of 91.66%, with an area under the ROC curve of 0.9452, indicating the value of applying the algorithm, which is based on the human vision system, to assess the image quality of non-mydratric photography, especially for low-cost ophthalmological telemedicine applications.

Index Terms—Human visual system, machine learning, portable fundus photography, quality assessment.

I. INTRODUCTION

OPTHALMOLOGICAL early detection using retinal photography prevents both vision impairment and the consequences of untreated eye disease. However, the current

lack of medical resources in some areas of developing countries, like China, leads to unfavorable outcomes. Non-mydratric ocular fundus photography seems to be a promising solution, especially for retinal disease, when combined with telemedicine [1], [2], because it does not require pupil dilation and it can be done with a portable fundus camera. Portable digital fundus photography differs from traditional fundus photography because it is done with the camera fixed to operator's hands rather than being on a permanent fixture. However, such operating conditions can be vulnerable to problems with the quality of the digital retinal image, such as uneven luminance, fluctuations in focus, and patients' movements. Hence, evaluating the image quality of portable fundus camera imaging-systems is of great importance.

The evaluation of fundus image quality involves a computer-aided retinal image analysis-system that is designed to assist ophthalmologists to detect eye diseases [3], such as age-related macular degeneration [4], glaucoma [5], [6], and diabetic retinopathy [7]. The objective quality evaluation of fundus images, which plays a major role in automatically selecting diagnosis-accessible fundus images among the outputs of digital fundus photography, is a descendant of subjective quality evaluation. Subjective quality evaluation is performed by experienced ophthalmologists who grade the quality of fundus images by comparing differences in the images to be graded with excellent quality images, based on their prior knowledge of excellent image quality. Such prior knowledge is acquired either from the human visual system (HVS), which is a complex biological system [8], or from technical training in ophthalmic diagnosis. Based on their prior knowledge, ophthalmological experts can grade fundus image quality with confidence; however, their subjective quality evaluation is as laborious in practice as it is expensive and time-consuming. Objective fundus image quality assessment is aimed at providing ophthalmologists with an alternative grading procedure, which not only takes less time, but also has equivalent accuracy.

Research related to the objective assessment of fundus image quality has been conducted for decades. The methods proposed in these studies can be classified into two major categories: generic feature-based methods [9]–[11] and structural feature-based methods [12]–[14]. Generic feature-based methods deal with global distortions, such as uneven illumination, blurring effects from being out of focus, and low contrast. Lee and Wang [15] presented an explicit template that was mathematically approximated by a Gaussian model to extract images of desired quality from a set of images. Convolution of the template with the intensity histogram of a retinal image was computed as generic quality. Fasih *et al.* [11] developed

Manuscript received September 24, 2015; revised November 30, 2015; accepted December 05, 2015. Date of publication December 08, 2015; date of current version March 31, 2016. This work was financially supported by the National Natural Science Foundation of China (81471748/H1812). S. Wang and K. Jin are joint first authors. *Asterisk indicates corresponding author.*

S. Z. Wang is with the Institute of VLSI Design, Zhejiang University, Hangzhou 310007, China (e-mail: longma624@163.com).

K. Jin is with the Department of Ophthalmology, the Second Affiliated Hospital of Zhejiang University, College of Medicine, Hangzhou 310009, China (e-mail: kingjinkai@126.com).

H. T. Lu is with the Institute of Translational Medicine, College of Medicine, Zhejiang University, Hangzhou 310020, China (e-mail: haitonglu6@126.com).

C. M. Cheng is with Med-imaging Integrated Solution Inc., Hsinchu Science Park, Hsinchu 30075, Taiwan.

*J. Ye is with the Department of Ophthalmology, the Second Affiliated Hospital of Zhejiang University, College of Medicine, Hangzhou 310009, China (e-mail: yejuan@zju.edu.cn).

*D. H. Qian is with the Institute of Translational Medicine, College of Medicine, Zhejiang University, Hangzhou 310020, China (e-mail: dqian@zju.edu.cn).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMI.2015.2506902

a generic retinal image quality estimation system that employed just noticeable blur (JNB), an HVS characteristic [16], combined with texture features [17]. Instead of finding an explicit template, implicit templates were utilized later on [10], including machine learning techniques such as the distance threshold of the k-Nearest Neighbours Classifier [18], support vectors embedded in a Support Vector Machine (SVM) [19], or weights in a Neural Network [20].

The HVS has many characteristics described by mathematics that humans can recognize as various patterns, such as colour [21], orientation [22], contour [23], motion [24], and frequency variation [25]. When applied to retinal image quality assessment, low-level characteristics of the HVS can extract generic features, such as illumination and colour, while high-level characteristics can extract structural features, such as vessel edges and macular texture. Here, we apply low-level HVS characteristics to generic quality assessment and propose an integrated HVS-based generic quality assessment algorithm as a starting point. Generic quality involves three parameters: illumination and colour, focus, and contrast. Three low-level characteristics of the HVS, including multi-channel sensation, just noticeable blur, the contrast sensitivity function, were employed to evaluate the three parameters, respectively.

The rest of the paper is organized as follows: Section II introduces materials containing public and proprietary retinal datasets and the subjective evaluation of images in these datasets; Section III describes the proposed algorithm, which consists of three parts: image preprocessing, HVS-based feature extraction, and a machine learning procedure; Section IV presents the results of tests of the algorithm on proprietary and public datasets, and compares these results with ophthalmologists' subjective evaluations; and Sections V and VI present the discussion and conclusion, respectively.

II. MATERIALS

Because the HVS-based algorithm needs to be trained on images with quality labels before predicting the generic quality of test images, the volume of the image datasets should be as large as possible. To collect our proprietary non-mydratic image datasets, a portable fundus camera was used, which was provided by the Med-imaging Integrated Solution Inc., Taiwan: the product type was DEC200. Due to the limited sample of patients and the scarcity of portable fundus cameras, there is a question whether that proprietary datasets represent the variety of generic quality distortion. Hence, some public retinal image datasets were added. Detailed information about both the proprietary and public datasets will be given first, followed by a subjective evaluation of them. Fig. 1 illustrates an example of a DEC200 portable fundus camera (a), two fundus images from the proprietary datasets (b), and two samples from public datasets (c).

A. Proprietary and Public Retinal Image Datasets

A total of 536 images were acquired to form four image datasets: two proprietary and two public datasets. Details are listed as follows:

Proprietary dataset #1 (LOCAL1): 199 images were captured by the DEC200 portable fundus camera. All images

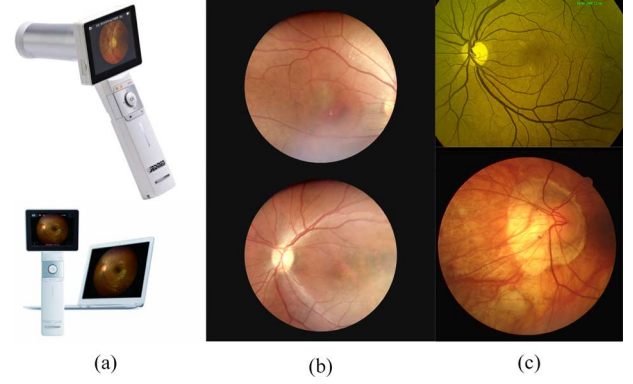


Fig. 1. Illustration of material samples: the DEC200 portable fundus camera (a); two images from proprietary datasets (b); two images from public datasets (c).

TABLE I
FOUR DATASETS INFORMATION

Dataset Name	FOV	Resolution	Number
LOCAL1	60°	2560 × 1960	199
LOCAL2	60°	2560 × 1960	103
DRIMDB [12]	60°	570 × 760	194
DRIVE [26]	45°	565 × 584	40

have a 600 field of view (FOV) and a resolution of 2560 by 1960 pixels.

Proprietary dataset #2 (LOCAL2): 103 images were captured by the same portable fundus camera, but during a different time period. All the images in LOCAL2 have the same FOV and resolution as in LOCAL1.

Public dataset #1 (DRIMDB): there were 216 images with 600 FOV and at 570 by 760 pixels. The original diabetic retinopathy image database (DRIMDB) consists of good (125 images), bad (69 images), outlier (22 images) categories according to the researchers' image quality grading system. In this paper, only the good and bad categories were used and the outlier class was excluded because of problems with the outlier image content. Although the images are from a diabetic retinopathy screening program, the generic-but-not-structural principle states that low generic quality must be distinguished from retinal abnormalities, like hard exudates, cotton wool spots, micro-aneurysms, and hemorrhages. The retinal abnormalities refer to structural quality which currently is not the topic of this paper. Public dataset #2 (DRIVE): there were 40 images with 450 FOV and a resolution of 565 by 584 pixels. Although this database is related to diabetic retinopathy, the subjective generic quality should be distinguished from retinal abnormalities based on the generic-not-structural principle.

Table I summarizes the key points of each dataset.

B. Subjective Evaluation

The work of grouping was performed according to the subjective evaluation of three ophthalmologists, one senior (marked by S) and two juniors (marked by J1, J2), who screened the images using a 0.275 mm per pixel monitor at a viewing distance of about 30 cm. Before inviting them to grade the total of 536 images, a generic quality gradation scale was made,

TABLE II
GENERIC QUALITY GRADATION SCALE

Grading bits	Principle description
0/1 --	Image is with the uneven illumination or colour
-- 0/1	Image has noticeable blur appeared in optic disks, vessels, or background, etc.
-- 0/1	Images contrast is low such that band of pixel intensity is narrow.

TABLE III
CONSISTENCY MATRIX OF INTER-OBSERVER GRADING

SROCC between	S and J1	S and J2	J1 and J2
Grading bit 1 (0/1 --)	0.6531	0.5853	0.6812
Grading bit 2 (-- 0/1)	0.6967	0.8016	0.7074
Grading bit 3 (-- 0/1)	0.5517	0.5805	0.4703

which adhered to the generic-but-not-structural principle, as listed in Table II. This scale is different from other schemes, like the Fleming's image clarity grading scheme [26], which mainly concerns small vessel clarity, and the field definition of the optic disk and macular.

The scale has three components that form a three-bit binary number. Therefore, the grading levels can be enumerated as the following: 000, 001, 010, 100, 101, 110, 111. For each of the three elements/digits, a one denotes that the principle description agrees with the image, whereas a zero denotes the opposite. Each image from all of the four datasets is associated with a three-digit binary subjective score, which indicates whether the image suffers from the uneven illumination or colour (I/C, indicated by the left digit or bit), noticeable blur distortion (Blur, indicated by the middle bit), and low contrast distortion (LC, indicated by the right bit).

As there were three ophthalmologists grading the subjective generic quality, the consistency of the inter-observer grading was evaluated. The Spearman rank order correlation coefficient (SROCC) was chosen as the consistency indicator, which is widely used in statistical analysis. For each grading bit in Table II, the SROCC between S and J1, S and J2, J1 and J2 were calculated. As there are three bits, a 3 by 3 inter-observer SROCC consistency matrix is shown in Table III.

C. Generic Quality Targets

To provide the reliable generic quality targets used for machine learning, subjective grades by the three ophthalmologists had to be pooled into golden-standard-like grades. For each of the three grading bits, as the inter-ophthalmologist SROCC analysis revealed that S's judgments were no better than J1's and J2's judgments, the majority voting rule was applied to each of the grading bits, such that if two or three observers agreed that an image had no partial distortion (assigning the responding grading bit with zero), then the image was labeled as having no partial distortion.

When the partial quality targets are determined using the majority voting rule, the overall generic quality targets can be acquired because good overall generic quality refers to not having any partial quality distortions. Hence, the partial quality targets of 000 are deemed to have good generic quality while the other targets have bad generic quality.

TABLE IV
DISTRIBUTION OF SUBJECTIVE TARGETS ON THE FOUR DATASETS

	I/C		Blur		LC		Overall	
	0	1	0	1	0	1	0	1
LOCAL1	41	158	105	94	114	85	27	172
LOCAL2	25	78	44	59	62	41	12	91
DRIMDB	140	54	120	74	118	76	103	91
DRIVE	40	0	40	0	40	0	40	0
total	246	290	309	227	334	202	182	354

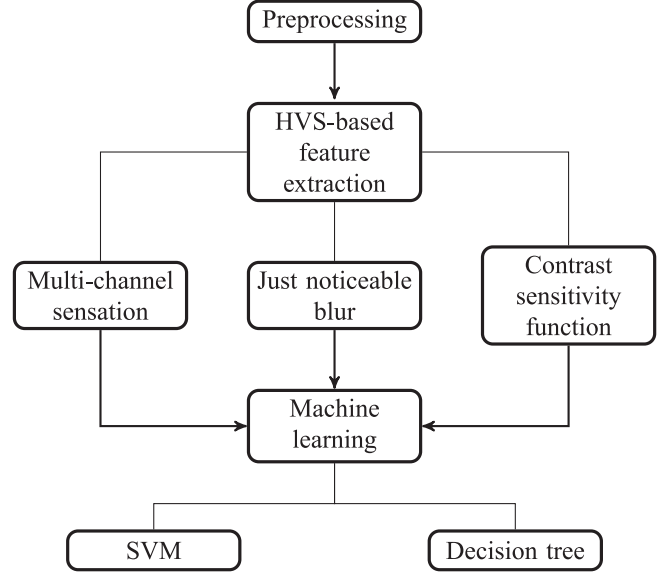


Fig. 2. The flow chart of the proposed algorithm.

As a summary on materials, the distribution of the three partial quality targets, as well as the overall quality targets, on the four datasets are shown in Table IV. The numbers refer to image volumes.

ALGORITHM

The major methodological steps presented here are outlined in the flowchart shown in Fig. 2. Irrelevant background was removed in the preprocessing step. The proposed HVS-based algorithm focuses on the feature extraction that is composed of three sub-models: multi-channel sensation, just noticeable blur, and contrast sensitivity function. The following machine learning step was devoted to evaluating the algorithm's capability of binary classification of images. Two machine learning tools, the SVM and the decision tree (DT) were employed.

D. Preprocessing

Preprocessing was designed to trim redundant background from the original retinal images. In order to obtain the trimming landmark, or mask, we combined boundary detection, which was implemented by Canny edge detection, and background thresholding. The boundary detection was aimed at detecting the furthest edge inside the foreground area and drawing a circular mask, with the center at the image center and the radius at the furthest edge. The background area was cropped and removed by this circular mask.

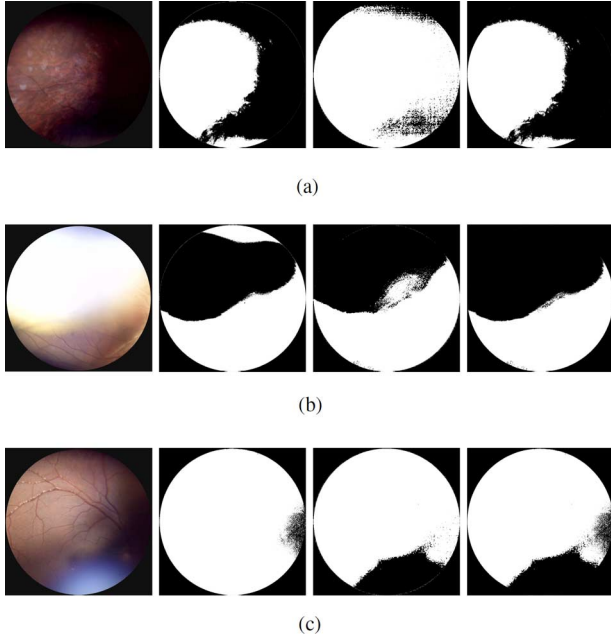


Fig. 3. Multi-channel distortion examples: each row represents one particular distortion type; from left to right: the RGB image, illumination mask, hue mask, and the final mask, respectively. (a) Low illumination distortion (b) High illumination distortion (c) Chromatic distortion.

E. HVS-Based Feature Extraction

Here we used HVS characteristics, including multi-channel sensation, just noticeable blur, and the contrast sensitivity function, to extract partial quality features. Multi-channel sensation deals with two visual channels, i.e., the illumination channel and the colour channel [28], both of which were first isolated from each other and then were processed to be an illumination mask and a colour mask, respectively. Finally, the two masks were integrated to represent perceived illumination-colour quality. Just noticeable blur is related to focus fluctuation. The term “just noticeable” means that the HVS can only detect the degree of blur distortion above a specific threshold and that any blur below the threshold fails to be perceived [29]. The contrast sensitivity function was based on previous experiments, which have indicated that the HVS perception of contrast correlates with spatial frequency [30]. Details about these three HVS-based techniques are elaborated in the following sections.

1) *Multi-Channel Sensation*: Multi-channel sensation is said to perceive illumination and colour distortion, examples of which are illustrated in Fig. 3. Illumination suffers from distortion in two ways: when illumination is too low (Fig. 3(a)) and when illumination too high (Fig. 3(b)); colour distortion refers to unsatisfactory colour, like that shown in Fig. 3(c). To detect all three types of problems, we transformed the RGB (R: red, G: green, B: blue) channels into HIS (H: hue, I: intensity, S: saturation) channels using a colour space transformation matrix:

$$\begin{pmatrix} I \\ H \\ S \end{pmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.274 & -0.322 \\ 0.211 & -0.523 & 0.312 \end{bmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \quad (1)$$

Here we use $\{I(x, y) | 1 \leq x \leq M, 1 \leq y \leq N\}$ to denote the pixel set of the intensity component whose size is M by N . Similarly, the previously preprocessed image mask can be expressed as $\{MI(x, y) | 1 \leq x \leq M, 1 \leq y \leq N\}$. Note that $MI(x, y)$ is a binary image, i.e., $MI(x, y) \in \{0, 1\}$.

The illumination mask was calculated by dual thresholds, T_{low} and T_{high} . The T_{low} was given by

$$T_{low} = \frac{1}{n} \sum_{n=1}^N \{I(x_i, y_i) | M(x_i, y_i) = 0\} \quad (2)$$

where n is total number of the pixels at which $M(x, y) = 0$.

The T_{high} was related to the T_{low} such that

$$T_{high} = 255 - \frac{T_{low}}{4}. \quad (3)$$

Now the illumination mask MI_1 can be obtained as

$$MI_1(x, y) = \begin{cases} 1, & T_{low} \leq I(x, y) \leq T_{high}, \\ 0, & \text{else.} \end{cases} \quad (4)$$

Similarly, the colour mask MI_2 was obtained using single thresholding on the hue component H which was obtained by colour space transformation seen in (1). The threshold T_{colour} was defined as

$$T_{colour} = \frac{1}{n} \sum_{i=1}^n \{H(x_i, y_i) | M(x_i, y_i) = 0\} \quad (5)$$

where the $H(x, y)$ denoted hue value at (x, y) coordinate. And the colour distortion mask was a binary image where

$$MI_2(x, y) = \begin{cases} 1, & H(x, y) > T_{colour}, \\ 0 & \text{else.} \end{cases} \quad (6)$$

The mask taking both illumination and colour distortion into consideration, denoted as MI_{ROI} , was obtained by applying logical AND operation on the M_1 and the M_2 :

$$MI_{ROI}(x, y) = \begin{cases} 1, & MI_1(x, y) = 1 \ \& \ MI_2(x, y) = 1 \\ 0, & \text{else.} \end{cases} \quad (7)$$

Multi-channel feature f_1 refers to the area ratio of MI_{ROI} to MI :

$$f_1 = \frac{\sum_{x,y=1}^{M,N} MI_{ROI}(x, y)}{\sum_{x,y=1}^{M,N} MI(x, y)}. \quad (8)$$

2) *Just Noticeable Blur*: Blur detection uses Cumulative Probability Blur Detection (CPBD) combined with the JNB, which have been applied to assess natural image quality [16] and to evaluate retinal image sharpness [11]. Both methods use Sobel edge detection to find an edge and then calculate the edge width. However, edge detection might introduce noise and the calculation of edge width is time-consuming. Here we combined the JNB detection method with a vessel density map because vessel density is vulnerable to blur. The vessel density map was obtained using a mathematical morphology algorithm [31]. To restrain noise, only vessel densities above 50 were retained and we denoted these vessel densities as the set $\{d(i)\}$.

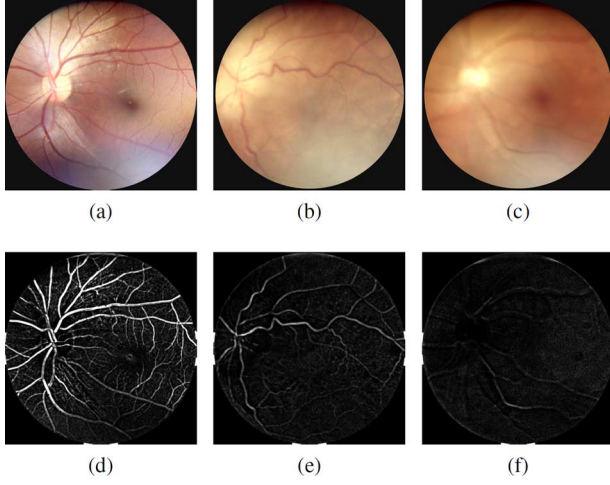


Fig. 4. Blur distortion examples: (a)-(c), retinal images with increasing blur distortion degree; (d)-(f) are the vessel density maps of (a)-(c), respectively.

Similar to the CPBD algorithm [11], the JNB vessel density d_{JNB} depends on the standard deviation (std) of $\{d(i)\}$:

$$d_{JNB} = \begin{cases} 110, & std(\{d(i)\}) \leq 13, \\ 120, & \text{else.} \end{cases} \quad (9)$$

The probability of just noticeable vessel density is expressed by:

$$P_{d(i)} = 1 - \exp\left(-\left|\frac{d(i)}{d_{JNB}}\right|\right). \quad (10)$$

The JNB probability P_{JNB} was defined as $d(i) = d_{JNB}$, which meant $P_{JNB} = 63\%$. The JNB feature f_2 was extracted from a normalized histogram of $P_{d(i)}$ such that

$$f_2 = P(P_{d(i)} \leq P_{JNB}) = \sum_{P_{JNB}}^1 P(P_{d(i)}). \quad (11)$$

As an illustration, RGB images with different degrees of blur distortion as well as vessel density maps are displayed in Fig. 4. Accordingly, Fig. 5 visualizes the probability histograms of $P_{d(i)}$ that were previously defined in (10). The JNB regions related to the JNB feature f_2 are marked with green. It can be observed that as the degree of blur distortion increases, the JNB regions shrink, indicating the JNB feature f_2 is competent to detect blur distortion.

3) *Contrast Sensitivity Function*: The workflow of using the contrast sensitivity function (CSF) to predict image quality has matured considerably over the last decade [30], [32]. Generally, there are three major steps for CSF-based image quality assessment.

- 1) Use the Fast Fourier Transform (*FFT*) to transform an image into a spatial frequency domain;
- 2) Choose a specified CSF to weight the spatial frequency coefficients;
- 3) Use the Inverse Fast Fourier Transform (*IFFT*) to transform the CSF-weighted spatial frequency coefficients back into spatial domain.

Such steps can be expressed mathematically as one equation:

$$\hat{I}(x, y) = IFFT\{FFT\{I(x, y)\} \times CSF(r)\} \quad (12)$$

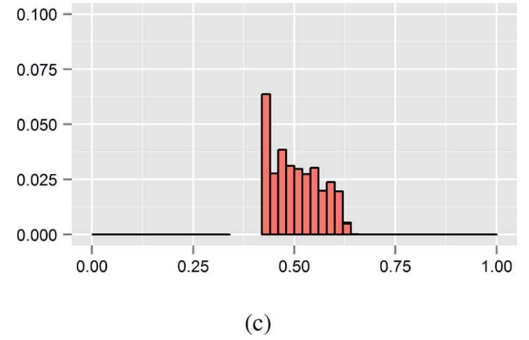
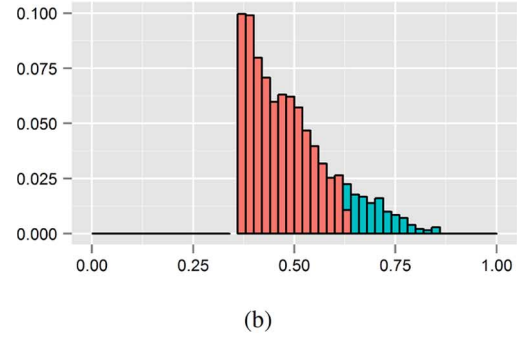
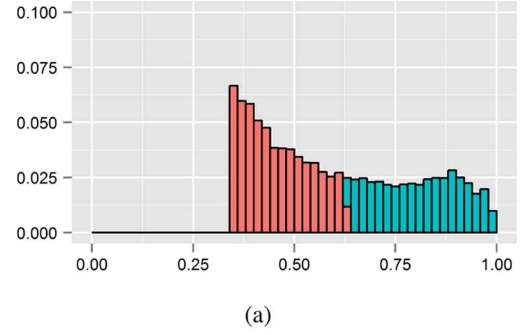


Fig. 5. Probability distribution histograms of vessel density maps, corresponding to retinal images shown in Figs. 4(a)–4(c): the JNB regions are coloured with green.

where the spatial frequency was denoted by r in polar coordinates and the symbol \times meant weighting by multiplication. The unit of the spatial frequency is circle per degree (cpd), which means that when viewed at the distance of 1 m, a 1.75 cm wide pattern will subtend an angle of 1 degree because the tangent of 1 degree is equal to 1.7455. Similarly, the maximum spatial frequencies associated with each dataset can be calculated precisely since the resolution of the images (seen in Table I), the physical pixel size (0.275 mm/pixel), and the viewing distance (30 cm) from the observer to the monitor are known. They are 71.3cpd, 71.3cpd, 41.0cpd, and 36.7cpd associated with LOCAL1, LOCAL2, DRIMDB, and DRIVE, respectively.

The $I(x, y)$ was the intensity image obtained from (1) and the $\hat{I}(x, y)$ was the output image of *IFFT*. As an illustration, Fig. 6 presents two images with low and high contrast in the left column.

The contrast sensitivity function $CSF(r)$ has mathematical models and Han and Cai [32] have provided a comprehensive

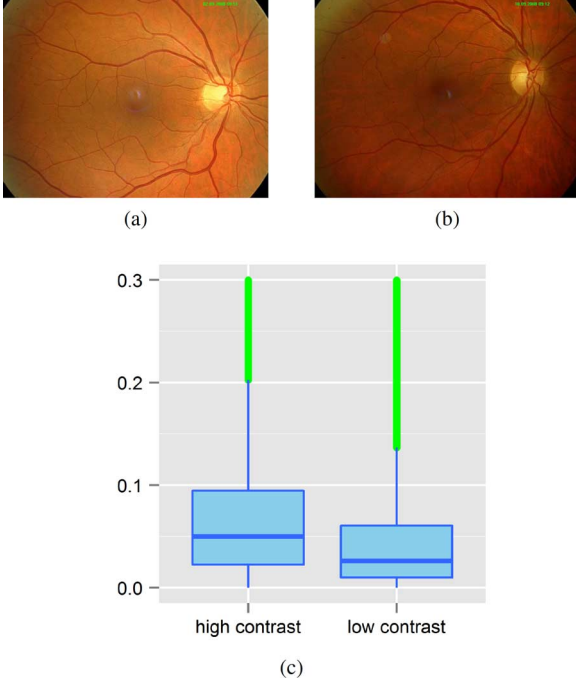


Fig. 6. Contrast distortion examples: (a) retinal image with high contrast, (b) retinal image with low contrast, (c) boxplot of the CSF-processed coefficients of (a) and (b).

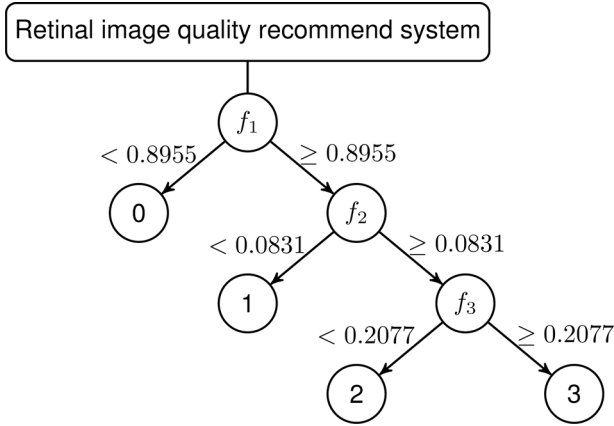


Fig. 7. Structure of the decision tree.

list of them. Considering the simplicity of model, we chose the CSF, which has the form of

$$CSF(r, a_0, a_1, a_2) = \left(1 - a_1 + \frac{r}{a_0}\right) \exp \left[- \left(\frac{f}{a_0} \right)^{a_2} \right] \quad (13)$$

where the parameters a_0, a_1, a_2 are such that $CSF(r)$ has a maximum value at $r = 11.3$ cpd [30]. Here the parameters are 11.3, 1, and 1, respectively. Note that the center spatial frequency does not stand outside the minimum band width because the minimum band width of the images in the four datasets is 36.7 cpd.

The output image $\hat{I}(x, y)$ was then normalized within 0 and 1 and the background pixels were removed by the ROI mask $MI_{ROI}(x, y)$.

The contrast sensitivity feature f_3 was computed according to the following equation:

$$f_3 = \frac{Q_3(\hat{I}) - Q_1(\hat{I})}{\max(\hat{I}) - \min(\hat{I})} \quad (14)$$

where Q_1 and Q_3 were the 1st and 3rd quartile values. The denominator referred to the range of $\hat{I}(x, y)$ with outlier values (drawn with green in Fig. 6) excluded.

F. Machine Learning

As HVS-based feature extraction generated three partial indicators of generic quality, i.e., color-illumination feature f_1 defined in (8), blur-detection feature f_2 defined in (11), contrast-detection feature f_3 defined in (14), the distortion-specified and the overall generic quality of a retinal image can be measured by the 3-dimensional feature space (f_1, f_2, f_3) . Here we use two representative types of machine learning algorithms: a SVM and a DT.

SVM is an excellent classification and regression tool, which uses support vectors that are trained from the training dataset to predict the testing dataset. It is a common practice to transform the original feature space into a higher dimensional space using kernel functions, such as a polynomial function, or a radial based function (RBF). In our experiment, the RBF kernel was adopted and relevant kernel parameters were automatically configured using the SVM toolbox of the KERNLAB package [33] written in R Language [34]. The binary classification problem, which needs to be resolved by SVM here, is used to predict the quality label (positive or negative) of retinal images based on the 3-dimensional feature vector (f_1, f_2, f_3) . The training and testing datasets were obtained by randomly dividing all 536 images from the four datasets described in Section II into two groups. To perform statistical testing, we did the random division 1000 times.

Rather than regarding the quality assessment of retinal images as a classification problem and solving it by SVM, we can also recommend, or select, retinal images whose generic quality is excellent. The recommend-system differs from classification in that each retinal image will have its own recommended index, not a class label. Retinal images with a recommended index above a determined threshold will be recommended, or selected. Here four recommend indices are defined and enumerated as 0, 1, 2 (bad generic quality) and 3 (good generic quality). DT is used to determine an appropriate recommendation index for each retinal image. The structure of our DT is outlined in Fig. 7, where the parameters are determined by threshold optimization, which will be described in next section.

III. RESULTS

This section illustrates the results of the quality assessment of retinal images. All these images have been graded by the three ophthalmologists, such that each retinal image has three partial quality grades and the overall quality grades: I/C (0: acceptable; 1: unacceptable), Blur (0: not noticeable; 1: noticeable), LC (0:

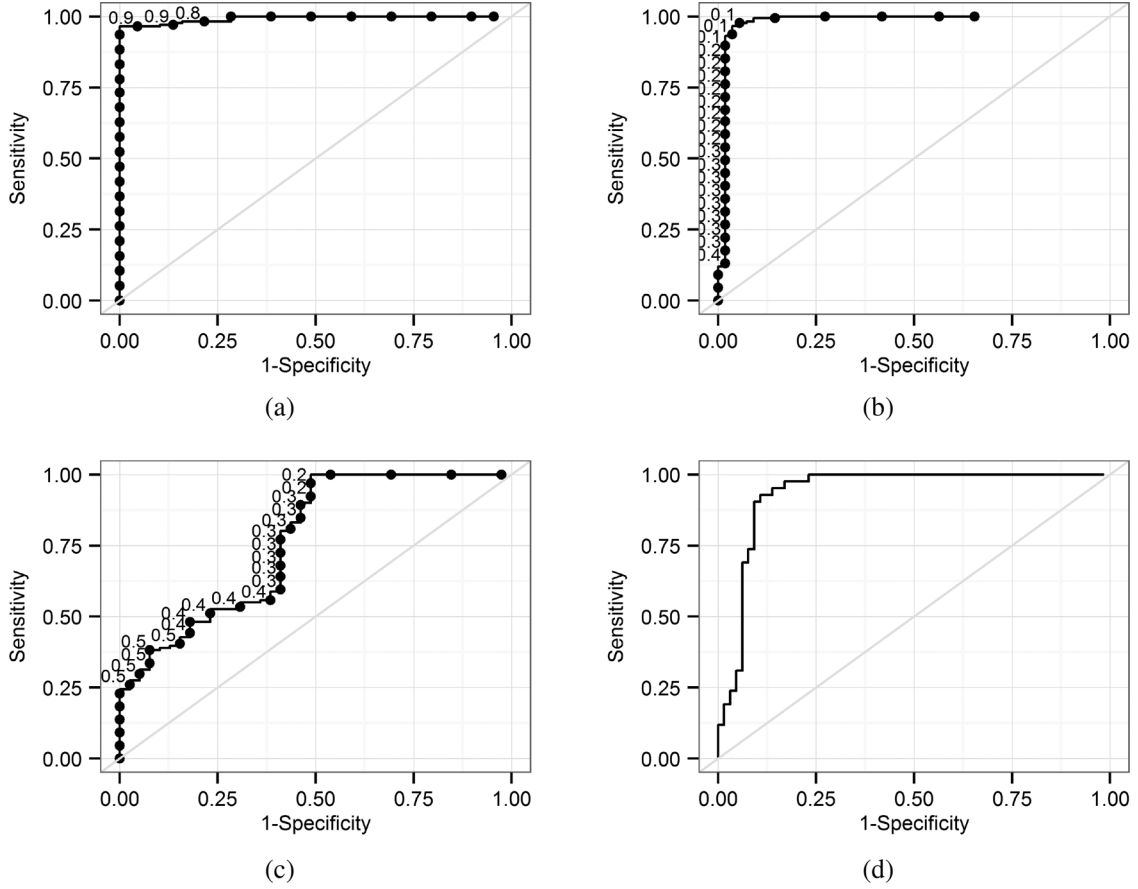


Fig. 8. ROC curves of: (a) illumination and colour classifier based on f_1 , (b) blur classifier based on f_2 , (c) contrast classifier based on f_3 , (d) generic quality classifier based on the feature vector f_1 , f_2 , f_3 . Numbers beside the points on each curve denote the classification threshold.

high contrast; 1: low contrast), and Overall (0: good generic quality; 1: bad generic quality).

Both the partial and overall quality assessments belong to a binary classification problem; the criteria commonly used in the evaluation of binary classifiers are the sensitivity and specificity defined by the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (15)$$

$$\text{specificity} = \frac{TN}{TN + FP}. \quad (16)$$

In our results, the positives mean that a retinal image has targets of zero. Accordingly, the three HVS-based features, f_1 , f_2 , f_3 , are formed as the feature vector to facilitate the quality prediction using SVM and DT.

A. Distortion-Specified Classification

The distortion-specified classification aims to detect images that suffer from specified generic degradations such as I/C, Blur, and LC. For each specified distortion, the training and testing set was collected according to the gold standard of subjective quality targets. The positive set consisted of images with the targets “000” and the negative set of I/C, Blur, and LC consisted of images with the targets “100”, “010”, and “001”, respectively.

Once the three sets were initially generated, the senior ophthalmologist checked them again and excluded the ambiguous images. Finally, the I/C set consisted of 261 images, with 172 positives and 89 negatives; the Blur set consisted of 176 positives and 56 negatives; the LC set consisted of 131 positives and 40 negatives.

The Receiver Operating Characteristic (ROC [35]) curves are plotted in Figs. 8(a)–8(c) to show the performance of the binary classifier based on f_1 , f_2 , f_3 , respectively. The numbers beside the points on each curve denote the classification threshold. The optimum thresholds are searched by finding the point having the shortest path towards the point where sensitivity = 1 and 1 – specificity = 0. Once the ROC curve is plotted, we can get sensitivity, specificity and the area under curve (AUC) associated with the optimum threshold. To perform statistical testing, we plotted the ROC curve plots 1000 times. Each plot was based on different subsets, with four fifth quantities of the whole set used for optimum thresholds searching and one fifth quantities used for performance measurement. Mean \pm Standard Deviation results of sensitivity, specificity, AUC, and the optimum thresholds are shown in Table V. The optimum thresholds were further used to construct a decision tree that classified the overall quality of all the images into two categories. The DT structure is outlined in Fig. 7. The performance parameters of the DT classification for overall generic quality are listed in Table V as well.

TABLE V
PERFORMANCE OF DISTORTION-SPECIFIED CLASSIFICATION BY SINGLE HVS-BASED FEATURE

Classifier	Category	Sensitivity	Specificity	AUC	Threshold
I/C	Acceptable	0.9658± 0.0286	0.9805± 0.0388	0.9732± 0.0240	0.8955± 0.0019
	Unacceptable	0.9805± 0.0388	0.9658± 0.0286		
Blur	Not noticeable	0.9692± 0.0290	0.9416± 0.0737	0.9554± 0.0373	0.0831± 0.0048
	Noticeable	0.9416± 0.0737	0.9692± 0.0290		
LC	High contrast	0.8174± 0.1046	0.5434± 0.1679	0.6799± 0.0844	0.2077± 0.0186
	Low contrast	0.5434± 0.1679	0.8174± 0.1046		
DT-based Overall	Good	0.8132	0.7712	N/A	N/A
	Bad	0.7712	0.8132		

TABLE VI
FEATURE VECTOR-BASED PERFORMANCE OF PARTIAL (I/C, BLUR, AND LC) AND OVERALL QUALITY CLASSIFIER USING SVM

SVM Classifier	Category	Sensitivity	Specificity	AUC
I/C	Acceptable	0.8797± 0.0746	0.8609± 0.0530	0.9310± 0.0239
	Unacceptable	0.8609± 0.0530	0.8797± 0.0746	
Blur	Not noticeable	0.8969± 0.0513	0.8815± 0.0476	0.9314± 0.0260
	Noticeable	0.8815± 0.0476	0.8969± 0.0513	
LC	High contrast	0.8776± 0.0515	0.8121± 0.0774	0.8783± 0.0354
	Low contrast	0.8121± 0.0774	0.8776± 0.0515	
SVM-based Overall	Good	0.8745± 0.0785	0.9166± 0.0352	0.9452± 0.0206
	Bad	0.9166± 0.0352	0.8745± 0.0785	

TABLE VII
PERFORMANCE OF CROSS-DATASET CLASSIFICATION ON ILLUMINATION AND COLOUR (I/C), BLUR (BLUR), AND LOW CONTRAST (LC)

SVM testing	DRIMDB			LOCAL1			LOCAL2		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
I/C	0.78± 0.03	0.74± 0.03	0.79± 0.00	0.68± 0.05	0.81± 0.03	0.78± 0.01	0.70± 0.08	0.89± 0.03	0.90± 0.00
Blur	0.89± 0.10	0.82± 0.10	0.93± 0.05	0.83± 0.06	0.73± 0.05	0.83± 0.01	0.83± 0.04	0.85± 0.03	0.83± 0.01
LC	0.84± 0.06	0.94± 0.13	0.95± 0.00	0.84± 0.06	0.55± 0.13	0.71± 0.01	0.84± 0.03	0.69± 0.04	0.79± 0.01
overall	0.87± 0.15	0.85± 0.03	0.90± 0.02	0.69± 0.18	0.80± 0.06	0.81± 0.01	0.68± 0.08	0.94± 0.02	0.94± 0.01

B. Partial and Overall Quality Classification Using SVM

Considering that some images have a combination of generic degradations, it was necessary to test the validation of the HVS-based feature vector, f_1 , f_2 , f_3 on the three types of partial quality (I/L, Blur, LC) and the generic overall quality; therefore, the SVM was used. The training and the testing sets were created using 5-fold cross-validation, with 80% images in the training set and 20% in the testing set. To get the statistical performance parameters, the cross-validation was repeated 1000 times. Each time the training and testing set was selected randomly. The Mean ± Standard Deviation results of sensitivity, specificity, and AUC are shown in Table VI.

C. Cross-Dataset Classification

Since different datasets have different resolution and the public datasets (DRIMDB, DRIVE) were not acquired by a DEC200 portable fundus camera, it was worthwhile to investigate the independence of the HVS-based feature vector of particular datasets. This was done by holding one dataset aside and doing the training with the other three datasets. The withheld dataset was used as the test set. As the DRIVE consisted of only good quality samples and the volume of DRIVE was relatively small, it was not included in the test sets. Therefore, the DRIMDB, LOCAL1, and LOCAL2 were tested based on the training sets of the other three datasets, respectively. Four quality indicators, three partial quality indicators (I/C, Blur, LC), and the generic overall quality

indicator (Overall) were investigated with each test dataset. Table VII summarizes the performance parameters, including the Mean ± Standard Deviation of the results of sensitivity, specificity, and AUC.

D. Noise Artifacts Influence

Modern complementary metal-oxide semi-conductor digital imaging system utilizes extensive image signal pre-processing algorithms to correct most of the noise sources induced by optics and image sensor, such as Gaussian noise, fixed pattern noise, and artifacts caused by dust. That is why retinal photography rarely suffers from noise and why images in this work do not encompass noise artifacts. Even so, for the sake of the HVS-based algorithm's competence for portable fundus camera application, it was decided to deliberately create a noise-contained dataset from 182 images which were labeled as having good generic quality. Three types of noise were used: Gaussian white noise (Gauss) with a mean of zero and a variance of 0.01, salt and pepper noise (Salt) with an intensity of 0.05, speckle noise (Speckle) with a mean of 0 and a variance of 0.04. For each type of noise, a SVM-based binary classifier was trained and tested on 364 images (182 noise-free and 182 with noise). Statistical performance parameters were obtained using a 5-fold cross-validation method that was run 1000 times, shown in Table VIII.

IV. DISCUSSION

In this study, three kinds of HVS characteristics were utilized to assess generic retinal image quality. Generic quality was

TABLE VIII
GENERIC QUALITY CLASSIFICATION ON NOISE ARTIFACTS

	Sensitivity	Specificity
Gauss($\mu = 0, \sigma = 0.01$)	100% \pm 0	100% \pm 0
Salt(intensity=0.05)	100% \pm 0	100% \pm 0
Speckle($\mu = 0, \sigma = 0.04$)	100% \pm 0	100% \pm 0

TABLE IX
SPEARMAN RANK ORDER CORRELATION COEFFICIENTS (SROCC)
MATRIX OF INTRA-OBSERVER GRADING

SROCC between	bit 1 and bit 2	bit 1 and bit 3	bit 2 and bit 3
S	0.3335	0.3726	0.5727
J1	0.2659	0.2579	0.5716
J2	0.3860	0.4906	0.6443

defined by the combination of three partial quality indicators: illumination-colour, blur, and contrast.

To detect the I/C distortion, our HVS-based method integrates illumination and colour into one single feature, rather than depicting the variance of the respective features [10]. For blur detection, there are two differences between our JNB-based method and other studies: one is that the structural-specified retinal vessel intensity was calculated to be the map associated with JNB rather than utilizing edge widths to quantify the threshold of JNB [11], [16]; on the other hand, other studies have added the probabilities of values below the edge width threshold, whereas the proposed JNB algorithm accumulates probabilities of vessel intensity above the threshold. In fact, both approaches are reasonable because retinal images suffering from blur distortion are less likely to having narrow edges or high vessel intensity than images with no blur distortion.

The distortion-specified classification based on individual HVS-based feature showed unbalanced performance: The I/C classifier based on the multi-channel sensation feature showed a classification sensitivity of 96.58% with a specificity of 98.05% while the Blur classifier based on the JNB feature had a sensitivity of 96.92% and a specificity of 94.16%. However, the contrast classifier based on the CSF feature showed a striking difference: a sensitivity of 81.74% and a specificity of only 54.34%, which caused us to wonder whether the contrast distortion had the relationship with I/C and Blur distortion from the perspective of HVS. To investigate this relationship, the consistency of the intra-ophthalmologist data was analysed. For each of the three observers, the SROCCs between each couple of three grading elements were calculated; Table IX lists the intra-observer consistency.

The intra-observer SROCCs between bit1 (I/C) and bit3 (LC), and bit2 (Blur) and bit3 show that low contrast is probably sensed by HVS as a mixture of the I/C and Blur, especially blur distortion, because the SROCCs in the right column in Table IX are higher than those in the other columns. This observation motivates us to use the feature vector, not the individual feature scalar, to predict partial and generic overall distortion. As listed in the Table VI, the performance of the LC classification is now competitive with I/C and Blur. The reason why performance of classification on I/C and Blur in Table VI is inferior to that using an individual feature in Table V is that images with a combination of individual distortions were included in the feature vector-based training and testing set.

Generic overall quality assessment integrating three partial quality indicators was performed by two different machine learning methods, SVM and DT. As seen in Table VI, SVM-based classification showed both high sensitivity (87.45%) and specificity (91.66%) while the DT-based recommendation method has a sensitivity of 81.32% with a specificity of 77.12%. The lower performance of the DT-based recommendation method may be due to the linear combination of the three partial quality indicators. However, the generic overall quality classification could be non-linearly separable, rather than linearly separable. In contrast, SVM does not have such a drawback for it can transform data into higher dimensional spaces where the non-linearly separable problem can be resolved linearly.

Different resolutions of the datasets and other artifacts, especially noise, are the other considerations. To handle these problems, cross-dataset classification and add-noise simulation were tested; shown in Tables VII and VIII, respectively. Due to different composition of each dataset, the performance of classifications of partial quality varies. However, the SVM-based overall quality classification testing on DRIMDB, LOCAL1, and LOCAL2 achieved performance above the AUC of 0.90, 0.81, and 0.94, respectively, indicating the adaptation of the HVS-based algorithm to different pixel size. Note that the running time of the algorithm is proportional to the pixel size, which would be a bottleneck for our algorithm being embedded in the portable camera imaging system. Hardware accelerating techniques, such as parallel pipeline processing, can be a solution.

Noise artifact simulation was done for the sake of completeness because DEC200 portable fundus cameras have gone through standard color image testing to make sure all the noise sources have been compensated for. However, the noise simulation also proves the effectiveness of the HVS-based features for sensing noise contamination.

V. CONCLUSION

This paper aimed to assess the generic quality of retinal images, especially for portable fundus camera applications in non-mydratic ocular fundus photography. An algorithm is proposed to assess image quality using a human vision system-based evaluation of three partial quality factors: illumination and colour, focus, and contrast. Accordingly, three characteristics of HVS were utilized: multiple channel sensation, just noticeable blur, and the contrast sensitivity function. The sensitivity of the distortion-specified classification was 96.58%, 96.92%, and 85.74%, respectively. The intra-ophthalmologists correlation analysis shows that HVS tends to mix low contrast distortion with blur and illumination and colour distortion. Therefore, using the HVS-based feature vector to predict low contrast is reasonable and effective, as Table VI shows that the three partial quality classifications achieve balanced performance. Different pixel sizes and noise artifacts are the other two factors. The cross-dataset classification shows the adaptation of the HVS-based features at different resolutions, with the SVM-based overall quality classification AUC of more than 0.80. Especially, the HVS-based feature can detect typical noise contamination such as Gauss noise, salt and pepper noise, and speckle noise, with both a sensitivity and specificity of 100%.

Meanwhile, a retinal image quality recommendation-system implemented by a decision tree showed a sensitivity of 81.32%.

Further studies are being planned and conducted, including developing a HVS-based structural quality assessment algorithm, the real-time implementation of the algorithm embedded in portable fundus cameras, and expanding the use of the technique to telemedicine applications.

ACKNOWLEDGMENT

The authors would like to thank Doctor Zhaoan Su, Doctor Yufeng Xu, and Doctor Lixia Lou for evaluating the quality of all of the retinal images.

REFERENCES

- [1] L. L. Shi *et al.*, "Telemedicine for detecting diabetic retinopathy: A systematic review and meta-analysis," *Br. J. Ophthalmol.*, vol. 99, no. 6, pp. 823–831, 2015.
- [2] K. Yogesan *et al.*, "Evaluation of a portable fundus camera for use in the tele-ophthalmologic diagnosis of glaucoma," *J. Glaucoma*, vol. 8, no. 5, pp. 297–301, 1999.
- [3] M. R. K. Mookiah *et al.*, "Computer-aided diagnosis of diabetic retinopathy: A review," *Comput. Biol. Med.*, vol. 43, no. 12, pp. 2136–2155, 2013.
- [4] M. U. Akram, A. Tariq, S. A. Khan, and M. Y. Javed, "Automated detection of exudates and macula for grading of diabetic macular edema," *Comput. Meth. Prog. Biomed.*, vol. 114, no. 2, pp. 141–152, 2014.
- [5] J. Cheng *et al.*, "Superpixel classification based optic disc and optic cup segmentation for glaucoma screening," *IEEE Trans. Med. Imag.*, vol. 32, no. 6, pp. 1019–1032, Jun. 2013.
- [6] 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. Imag.*, vol. 30, no. 6, pp. 1192–1205, Jun. 2011.
- [7] C. Kose, U. Sevik, C. Ikibas, and H. Erdol, "Simple methods for segmentation and measurement of diabetic retinopathy lesions in retinal fundus images," *Comput. Meth. Prog. Biomed.*, vol. 107, no. 2, pp. 274–293, 2012.
- [8] Y. Adini, D. Sagi, and M. Tsodyks, "Context-enabled learning in the human visual system," *Nature*, vol. 415, no. 6873, p. 790, 2002.
- [9] A. G. Marrugo, M. S. Millan, M. Sorel, J. Kotera, and F. Sroubek, "Improving the blind restoration of retinal images by means of point-spread-function estimation assessment," in *Proc. SPIE*, 2015, vol. 9287, p. 6.
- [10] J. M. P. Dias, C. M. Oliveira, and L. da Silva Cruz, "Retinal image quality assessment using generic image quality indicators," *Inf. Fusion*, vol. 19, no. 1, pp. 73–90, 2014.
- [11] M. Fasih, J. M. P. Langlois, H. B. Tahar, and F. Cheriet, "Retinal image quality assessment using generic features," in *Med. Imag. 2014: Comput. Aid. Diag.*, 2014, vol. 9035.
- [12] U. Sevik, C. Kose, T. Berber, and H. Erdol, "Identification of suitable fundus images using automated quality assessment methods," *J. Biomed. Opt.*, vol. 19, no. 4, p. 046006, 2014.
- [13] J. Paulus, J. Meier, R. Bock, J. Hornegger, and G. Michelson, "Automated quality assessment of retinal fundus photos," *Int. J. Comput. Assist. Radiol.*, vol. 5, no. 6, pp. 557–564, 2010.
- [14] L. Giancardo *et al.*, "Elliptical LOCAL vessel density: a Fast and robust quality metric for retinal images," in *Proc. IEEE EMBC*, 2008, pp. 3534–3537.
- [15] S. C. Lee and Y. M. Wang, "Automatic retinal image quality assessment and enhancement," in *Proc. SPIE*, 1999, vol. 3661, pp. 1581–1590.
- [16] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the Cumulative probability of blur detection (CPBD)," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2678–2683, Sep. 2011.
- [17] Y. Ming-Hsuan, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, Jan. 2002.
- [18] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-Nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. 15, no. 4, pp. 580–585, Jul./Aug. 1985.
- [19] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [20] A. K. Jain, R. P. W. Duin, and J. C. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.
- [21] C. McCollou, "Color adaptation of edge detectors in human visual system," *Sci.*, vol. 149, no. 3688, p. 1115, 1965.
- [22] F. W. Campbell and J. J. Kulikows, "Orientational selectivity of human visual system," *J. Physiol. London*, vol. 187, no. 2, p. 437, 1966.
- [23] D. J. Field, A. Hayes, and R. F. Hess, "Contour integration by the human visual system evidence for a LOCAL association field," *Vis. Res.*, vol. 33, no. 2, pp. 173–193, 1993.
- [24] Z. L. Lu and G. Sperling, "Three-systems theory of human visual motion perception: review and update," *J. Opt. Soc. Am. A.*, vol. 18, no. 9, pp. 2331–2370, 2001.
- [25] C. Braccini, G. Gambardella, and G. Sandini, "A signal theory approach to the space and frequency variant filtering performed by the human visual system," *Signal Process.*, vol. 3, no. 3, pp. 231–240, 1981.
- [26] A. D. Fleming, S. Philip, K. A. Goatman, J. A. Olson, and P. F. Sharp, "Automated assessment of diabetic retinal image quality based on clarity and field definition," *Invest. Ophth. Vis. Sci.*, vol. 47, no. 3, pp. 1120–1125, 2006.
- [27] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.
- [28] A. Toet and M. P. Lucassen, "A new universal colour image fidelity metric," *Displays*, vol. 24, no. 4–5, pp. 197–207, 2003.
- [29] R. Ferzli and L. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 717–728, Apr. 2009.
- [30] G. Ginesu, F. Massidda, and D. D. Giusto, "A multi-factors approach for image quality assessment based on a human visual system model," *Signal Process. Image Commun.*, vol. 21, no. 4, pp. 316–333, 2006.
- [31] F. Zana and J. C. Klein, "Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1010–1019, Jul. 2001.
- [32] Y. Han and Y. Cai, "Contrast sensitivity function calibration based on image quality prediction," *Opt. Eng.*, vol. 53, no. 11, p. 113107, 2014.
- [33] K. A. Smola, K. Hornik, and A. Zeileis, "Kernlab: An S4 package for kernel methods in R," *J. Stat. Softw.* vol. 11, no. 9, pp. 1–20 [Online]. Available: <http://www.jstatsoft.org/v11/i09>
- [34] "Team R: a Language and environment for statistical computing," in *R Found. Stat. Comput.*, Vienna, Austria [Online]. Available: <http://www.R-project.org/>=0pt
- [35] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett. Team.*, vol. 27, no. 8, pp. 861–874, 2006.