Automated Quality Assessment of Retinal Fundus Photos

**Jan Paulus · Jörg Meier · Rüdiger Bock ·
Joachim Hornegger · Georg Michelson**

Received: date / Accepted: date

Abstract *Objective* Automated, objective and fast measurement of the image quality of single retinal fundus photos to allow a stable and reliable medical evaluation.

Methods The proposed technique maps diagnosis-relevant criteria inspired by diagnosis procedures based on the advise of an eye expert to quantitative and objective features related to image quality. Independent from segmentation methods it combines global clustering with local sharpness and texture features for classification.

Results On a test dataset of 301 retinal fundus images we evaluated our method on a given gold standard by human observers and compared it to a state of the art approach. An area under the ROC curve of 95.3% compared to 87.2% outperformed the state of the art approach. A significant p-value of 0.019 emphasizes the statistical difference of both approaches.

Conclusions The combination of local and global image statistics models the defined quality criteria and automatically produces reliable and objective results in determining the image quality of retinal fundus photos.

Keywords Retina · fundus image · quality assessment · non-reference image quality metric

1 Introduction

1.1 Motivation

Medical images are a very important basis for diagnosis and patient treatment. In particular in ophthalmology photos of the eye background are used by medical experts to diagnose and document diseases like glaucoma or diabetic retinopathy. In addition the images are commonly further evaluated by automatic software tools to support the diagnosis [1–3].

Sufficient image quality is essential to ensure a reliable diagnosis and a valid automated processing. Because of the operating personnel's varying level of experience, different types of cameras or the individual properties of the acquired eye the quality of images highly varies. Photos of poor quality should not be further used for diagnosis. A reacquisition would be necessary. However, in many cases like in reading centers in Germany and the USA [4], the image acquisition is time and location independent from its medical assessment. A reacquisition of the images will be time consuming and expensive. Thus a sufficient image quality has to be assured already during the acquisition procedure.

Unfortunately the rating of image quality is subjective and application dependent. It is an individual decision at which point the image quality becomes too bad for a stable diagnosis. There is a strong need to objectify image quality during the acquisition. This would help to ensure an overall sufficient quality level for the acquired image data that is essential for a stable and reliable diagnosis.

J. Paulus · J. Meier · R. Bock · J. Hornegger
Pattern Recognition Lab
Graduate School in Advanced Optical Technologies (SAOT)
Martensstr. 3, 91058 Erlangen, Germany
Friedrich-Alexander-University Erlangen-Nuremberg

G. Michelson
Department of Ophthalmology
Graduate School in Advanced Optical Technologies (SAOT)
Interdisciplinary Center of Ophthalmic Preventive Medicine and Imaging (IZPI)
Schwabachanlage 6, 91054 Erlangen, Germany
Friedrich-Alexander-University Erlangen-Nuremberg

1.2 State of the art

In literature, the main purpose for automated quality assessment in common images is to compare original images to their compressed versions for quality loss quantification, so called reference approaches. Eskicioglu et al. [5] provide an overview of basic quality metrics for this problem, such as average difference or normalized cross-correlation. Several works in that field develop extended approaches [6] e.g. driven by the human eye's function of finding structures [7].

In the field of medical imaging those reference approaches used for common images are not feasible as comparable reference images are rarely available. Despite of the importance of this problem it is still a widely neglected field of research especially with regard to ophthalmic fundus imaging. To the authors' knowledge there are only five relevant publications dealing with retinal image quality assessment: (i) Segmentation based approaches detect anatomical structures, while it is assumed that the segmentation will fail on low quality images due to the bad recognizability. Fleming et al. [8] measure the quality by evaluating the vessel tree in the region around the macula (point of sharpest vision in the retina). In addition, anatomical criteria related to the optic nerve head (exit of the optic nerve out of the retina) and the macula describe an image formation that is required to achieve good quality images. Giancardo et al. [9] measure the densities of vessels for different regions in the image. The vessel densities and a 5-bin-histogram of each color channel are used as features for classification. (ii) Histogram based approaches use information gained by image statistics to identify low quality photos. Lalonde et al. [10] evaluate the histogram of an input image's gradient magnitude image and local histogram information of its gray values. Reference histograms are calculated out of images showing good quality and compared with the input image's histograms for classification. Lee et al. [11] compute a quality index by convolving the intensity histogram of the input image with the template intensity histogram from good retinal images. *Image Structure Clustering* (ISC) [12] characterizes the image quality by the distribution of image intensities itself and the ability to cluster the image into the contained anatomical structures. Five clusters are calculated from the input image using a bank of filters to transform the pixels into the gauge coordinate system that is defined at each point by the direction of its gradients.

1.3 Contribution

Most of the state of the art methods focus either on segmentation methods, that can be error-prone, or on histogram information, that misses the structural information of relevant components. As an exception ISC incorporates the promising idea of assessing the structural recognizability of anatomical components but mainly uses local gradient information of a non-objective gold standard. We seize the idea but present a new method that introduces a combination of global and local structural characteristics as a non-reference approach and waives error-prone segmentation. In contrast to the state of the art it is driven by four criteria inspired by diagnosis procedures based on the advise of an eye expert. By judging an image according to these criteria quality assessment becomes a more objective task and enables the building of an objective gold standard (figure 1). The criteria are designed for the application on optic nerve head centered fundus images of 22.5° field of view. Anatomical components like the fovea are not visible and will not be considered in the following:

- **Structural criteria**
  1. *Optic disk structure*
     Can we recognize and differentiate the structure of the optic disk?
  2. *Vessel structure*
     Can we recognize and differentiate the fine structure of the vessels?
- **Generic criteria**
  3. *Homogeneous illumination*
     Is the illumination and brightness approximately equal in all parts of the image?
  4. *Bright and high-contrast background*
     Is the eye's background bright enough and of sufficient contrast?

The structural criteria are covered by an unsupervised **clustering** and a **sharpness** metric. Like in ISC the clustering groups the anatomical structures into clusters. ISC uses a bank of complex filters for a gauge coordinate transformation. Therefore, it mainly focuses on gradient and thus local information. In contrast, we gain global information using a more basic operation by applying k-means-clustering directly on the pixel intensities. We also utilize cluster sizes to express the size of relevant components. Another advantage of this basic operation is the possibility to compute inter-cluster-differences for the description of the recognizability and dissimilarity of these anatomical structures. Like ISC we incorporate local gradient information, but we gain it separately as the sharpness metric measures the clearness of separation between the components.

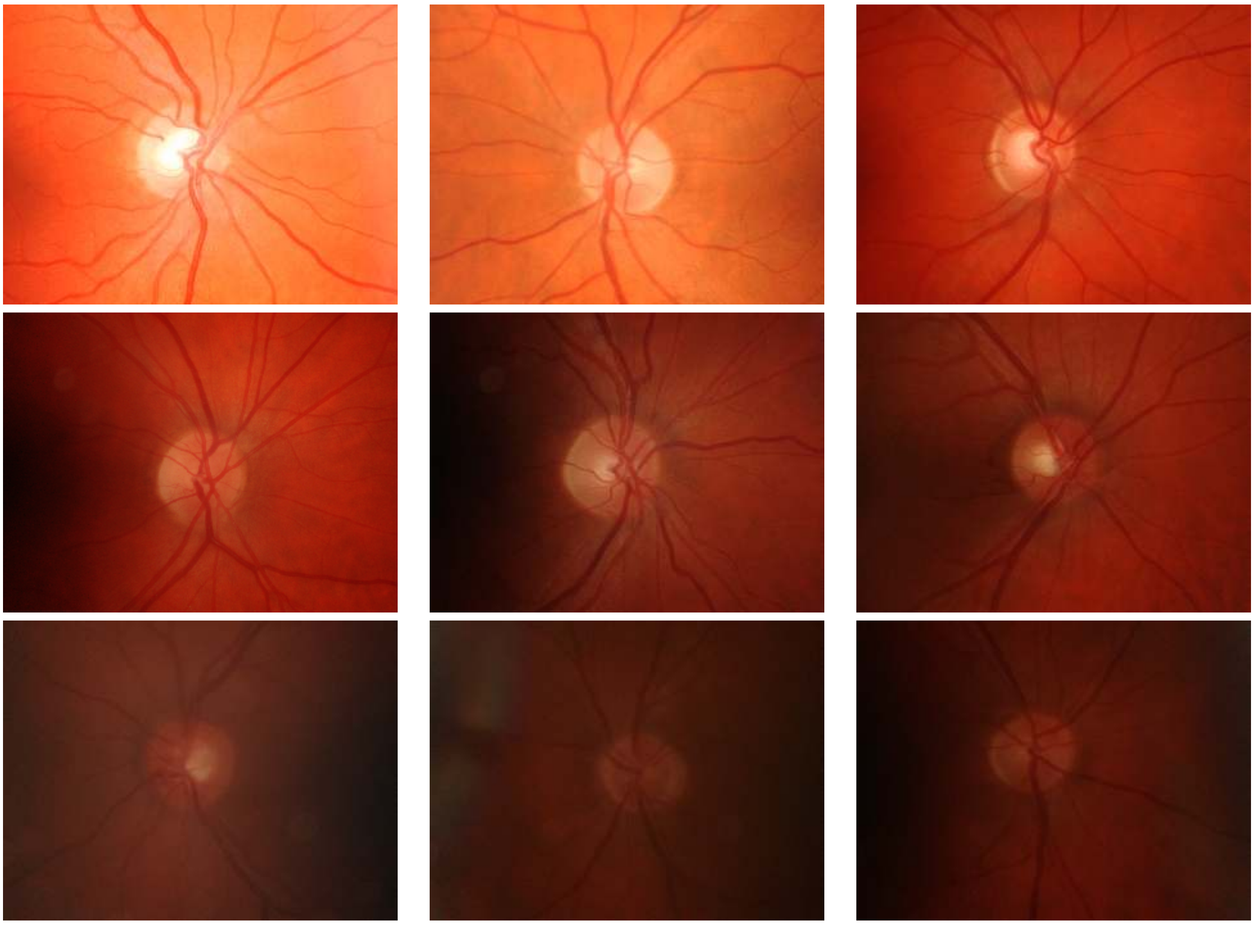


Fig. 1 Example for retinal fundus images of excellent (upper row, all criteria fulfilled), average (middle row, two criteria fulfilled) and insufficient (lower row, no criteria fulfilled) quality. The images of excellent quality show clearly the optic disk (bright circular spot in the middle where the optic nerve exits the eye, also known as "blind spot"), the vessel tree (exiting into the eye at the optic disk), a high-contrast background and an overall homogeneous illumination. The rating is based on the majority decision of three human evaluators using the criteria defined in section 1.3. Excellent and average quality will be considered to be sufficient for further use and referred as *good* quality. The average quality images show the problem of judging quality at the class border. Insufficient quality indicates a reacquisition and will be included in the set of *bad* quality images in the following.

As a major improvement we introduce the **Haralick** texture metrics [13] into the field of retinal quality to describe the generic criteria. Beside the sharpness of the image the Haralick metrics evaluate the homogeneity and the contrast.

Summarizing, the clustering describes the recognizability, dissimilarity and contrast of relevant structures. The sharpness metric evaluates the separation between components. The Haralick features measure common image sharpness, homogeneity and generic contrast. Thus we combine global and local information which is not yet present in this form in the state of the art.

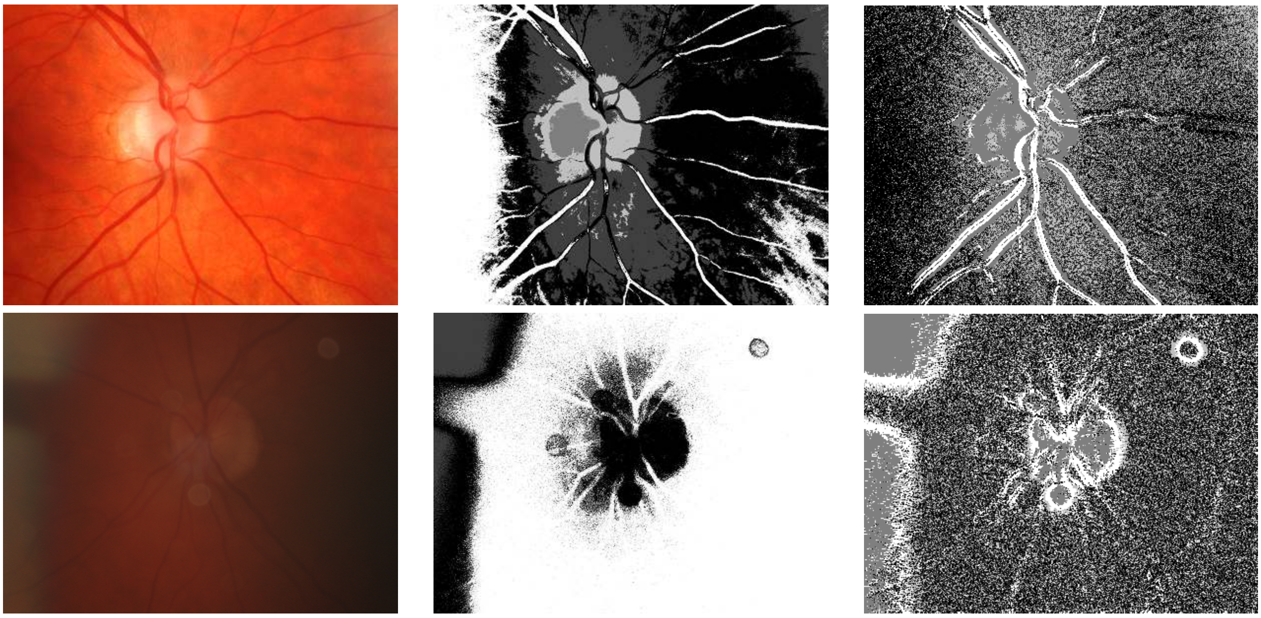
2 Methods

Our algorithm models the criteria defined above to measure the image quality that is relevant for a reliable assessment of fundus images. The method consists of a clustering, a sharpness metric and Haralick texture features.

We combine all features in one final vector. For all computations only the green channel was considered as it shows the best contrast.

2.1 Clustering

As we want to assure sufficient recognizability and differentiation of anatomical structures (e.g. optic disk, vessels) we identify these components by applying a k-means-clustering of the input image I of size $n \times m$ with k clusters C_i with $i \in \{1, \ldots, k\}$. The gray values g_{xy} with $x \in \{1, \ldots, n\}$ and $y \in \{1, \ldots, m\}$ are grouped in clusters without further preprocessing.



(a) Input image **(b)** Clustering result (proposed method) **(c)** Clustering result (ISC)

Fig. 2 Clustering examples: *Good* (first row) and *bad* (second row) quality images (a) and clustering results for the proposed method (b) and ISC as state of the art (c) coded as gray values. For *good* quality images the clustering images show the characteristic anatomical structures. In the case of *bad* quality they are not recognizable.

The cluster centers are initialized with mean values of the k structures (e.g. vessels) in 10 images manually segmented by one person. The images showed *good* quality and were considered by three human evaluators to fulfill all quality criteria. In each image representative pixels for each cluster were identified and their intensities averaged for each cluster over all 10 images.

In *good* quality images each anatomical structure has an expected size where significant variations refer to bad recognizability and thus *bad* quality. We assess the structure size by using the normalized cluster sizes c_i as features, where # denotes the cardinal number.

$$c_i = \frac{\#\{g_{xy} | g_{xy} \in C_i\}}{n \cdot m} \quad (1)$$

The clearer we can recognize certain structures and differentiate between them the higher their inter-cluster-contrast. We use inter-cluster-differences as essential features to express this structural contrast. They are generated by computing the difference d_{ij} between the mean value m_i of a certain cluster C_i and all other clusters' mean values m_j.

$$d_{ij} = m_i - m_j, i \in \{1, \ldots, k\}, j \in \{1, \ldots, k\}, i > j \quad (2)$$

Thus the cluster sizes c_i and the inter-cluster-differences d_{ij} evaluate the structural recognizability and dissimilarity of relevant image components like e.g. the optic disk. For *bad* quality images the clustering will consequently fail resulting in abnormal cluster sizes and low inter-cluster-differences (figure 2).

2.2 Sharpness

Our clustering (section 2.1) measures the differentiation of relevant structures globally. It does not cover local properties at the structures' borders where a clear and sharp edge is important for *good* quality as it will separate the components (e.g. optic disk, vessels) from each other more clearly. Therefore we incorporate a sharpness metric that evaluates the edge strength in the image. High gradients identifying sharp edges we calculate the gradient magnitude image G of the input image I by combining the derivative I_x in x-direction and the derivative I_y in y-direction using the Euclidean norm.

$$G = \sqrt{I_x^2 + I_y^2} \; with \; I_x = \frac{\partial I}{\partial x}, I_y = \frac{\partial I}{\partial y} \quad (3)$$

The gray values e_{xy} in the gradient magnitude image G are normalized to the range $[0; 1]$ by a minimum maximum scaling. We use the normalized number of pixels identifying strong edges s_1 and the average strength of strong edges s_2 to express the image sharpness. Strong edges have to lie above a threshold $\alpha \in [0; 1]$, that was empirically set to twice the mean gray value in G.

$$s_1 = \frac{\#\{e_{xy} | e_{xy} \geq \alpha\}}{n \cdot m} \tag{4}$$

$$s_2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} v_{ij}}{\#\{e_{xy} | e_{xy} \geq \alpha\}}, \; v_{ij} = \begin{cases} 0 & e_{ij} < \alpha \\ e_{ij} & e_{ij} \geq \alpha \end{cases} \tag{5}$$

$$\alpha = \frac{2 \sum_{i=1}^{n} \sum_{j=1}^{m} e_{ij}}{n \cdot m} \tag{6}$$

Thus both features s_1 and s_2 indicate how clearly the structures (e.g. optic disk, vessels) are separated from each other.

2.3 Haralick

To incorporate generic image quality statistics we compute three Haralick metrics [13] that are well known as texture metrics. We are using entropy h_1 as description for common image sharpness, energy h_2 as description for image homogeneity and contrast h_3. They are based on so called co-occurrence matrices which contain the relative frequencies $P(i, j, r)$ counting the neighborhood of each gray value to each other gray value in direction $r \in \{0°, 45°, 90°, 135°\}$.

$$\begin{aligned} P(i, j, 0°) = \#\{&(a, x) \in [1, \dots, n], (b, y) \in [1, \dots, m] \mid \\ &g_{ab} = i, \, g_{xy} = j, \, a - x = 0, \, |b - y| = 1\} \end{aligned} \tag{7}$$

$$\begin{aligned} P(i, j, 45°) = \#\{&(a, x) \in [1, \dots, n], (b, y) \in [1, \dots, m] \mid \\ &g_{ab} = i, \, g_{xy} = j, \, (a - x = 1, \, b - y = -1) \vee \\ &(a - x = -1, \, b - y = 1)\} \end{aligned} \tag{8}$$

$$\begin{aligned} P(i, j, 90°) = \#\{&(a, x) \in [1, \dots, n], (b, y) \in [1, \dots, m] \mid \\ &g_{ab} = i, \, g_{xy} = j, \, |a - x| = 1, \, b - y = 0\} \end{aligned} \tag{9}$$

$$\begin{aligned} P(i, j, 135°) = \#\{&(a, x) \in [1, \dots, n], (b, y) \in [1, \dots, m] \mid \\ &g_{ab} = i, \, g_{xy} = j, \, (a - x = 1, \, b - y = 1) \vee \\ &(a - x = -1, \, b - y = -1)\} \end{aligned} \tag{10}$$

Each matrix entry is normalized by the total number of neighbored pixel pairs in its certain direction N_r.

$$p(i, j, r) = \frac{P(i, j, r)}{N_r} \tag{11}$$

Based on the four co-occurrence matrices entropy h_1^r, energy h_2^r and contrast h_3^r are calculated for each direction r.

$$h_1^r = -\sum_{i=1}^{m \cdot n} \sum_{j=1}^{m \cdot n} p(i, j, r) log(p(i, j, r)) \tag{12}$$

$$h_2^r = \sum_{i=1}^{m \cdot n} \sum_{j=1}^{m \cdot n} p(i, j, r)^2 \tag{13}$$

$$h_3^r = \sum_{l=0}^{m \cdot n - 1} l^2 \{ \sum_{i=1}^{m \cdot n} \sum_{\substack{j=1 \\ |i-j|=l}}^{m \cdot n} p(i, j, r)\} \tag{14}$$

The final Haralick features h_1, h_2 and h_3 are generated by computing the mean of all directions.

$$h_1 = \frac{1}{4} \sum_r h_1^r \tag{15}$$

$$h_2 = \frac{1}{4} \sum_r h_2^r \tag{16}$$

$$h_3 = \frac{1}{4} \sum_r h_3^r \tag{17}$$

Thus, texture statistics are used to calculate generic quality features, entropy h_1 for common image sharpness, energy h_2 for image homogeneity and contrast h_3.

2.4 Feature Composition

The k cluster sizes c_i, the inter-cluster-differences d_{ij}, the two sharpness metrics s_1, s_2 and the Haralick features h_1, h_2 and h_3 are combined in one final feature vector. After evaluating the classification performance, we chose $k = 5$ for the clustering detecting two optic disk regions (cup and rim), two background regions (brighter and darker background) and the vessels. The gained 20-dimensional feature vector is directly used for classification.

3 Materials and Results

3.1 Materials

Our evaluation data set consisted of 301 retinal color fundus photos acquired with a Kowa non-myd camera. The images are optic disk centered and have a size of 1600 × 1212 pixels with a field of view (FOV) of 22.5°. The data set contained the 10 images used for the initialization of the clustering. Three human observers including one eye expert evaluated the data set. Each observer decided on the quality for each image using the criteria defined in section 2. An image was considered *good*, and thus sufficient for a reliable diagnosis fulfilling at least two criteria and *bad* otherwise. The inter-observer-correlation indicated by Fleiss' κ results into $\kappa = 0.58$. For each image the label classified by the majority of the three observers was defined as an overall quality gold standard. In this manner the data set was divided into 236 *good* and 65 *bad* fundus photos.

3.1.1 Experimental Setup

We evaluated the proposed method by testing all feature subsets and their possible combinations. The result of the combination of all features was compared to *Image Structure Clustering* (ISC). ISC is like the proposed method a non-segmentation based approach. It will be also referred as state of the art in the following.

In our implementation of ISC we applied two modifications divergent from [12]. (i) For speed improvements we halved the image size of each input image using subsampling. This seems valid, since in [12] the authors applied a resampling as well to gain an unique field of view out of two types of image sizes and fields of view. (ii) We initialized the mean vectors of each of the five clusters with the same manually segmented data used for the initialization of our clustering (section 2.1). For each pixel of each pre-segmented cluster the filter bank respond was calculated. The resulting vectors were average over all 10 images.

In each of our experiments we performed a 10-fold-cross-validation. Images were chosen randomly for each of the 10 subsets. Each image appeared exactly once in the experiment. Five subsets consisted of 6 *bad* and 24 *good* images, four subsets of 7 *bad* and 23 *good* images and one subset of 7 *bad* and 24 *good* images. Each subset was used for testing exactly once and the remaining folds for training the classifier.

3.1.2 Classifier Setup

We used a Support Vector Machine (SVM) with a radial basis function $k(\mathbf{u}, \mathbf{v}) = \exp(-\gamma \cdot |\mathbf{u} - \mathbf{v}|^2)$ as classifier in all experiments. The variance γ of the radial basis kernel and the penalty factor C were calculated using a grid search strategy in order to find the best parameter set for each method. For each parameter exists a particular value range that is applied in a certain step width. During a cross-validation process the parameter combination resulting into the lowest classification error is chosen for the evaluation. The parameters for ISC and for our proposed method and its feature subsets were applied on a libSVM [14] implementation during the whole evaluation.

3.2 Results

For quantifying the performance of the proposed method we calculated the area under the ROC curve (AUC), the p-value related to ISC and the p-value related to the final feature combination of Haralick, clustering and sharpness features. For computing the accuracy our probabilistic SVM used 50.0% as classification threshold. The sensitivity for classifying *bad* images at a specificity of 80.0% is given, since the *bad* images have to be identified to enable a reacquisition on the fly. The chosen threshold seems to be a sufficient value for detecting good images, since accepted *bad* images implicate higher costs. The results were computed for each feature combination of the proposed method. The performance of all features in every combination were compared to ISC as state of the art for non-segmentation based approaches (table 1).

The Haralick features reach the highest performance of the isolated features on our data set (90.7% sensitivity, 89.7% accuracy, 92.7% AUC) which increases by incorporating the clustering (93.9% sensitivity, 90.4% accuracy, 94.0% AUC) and the sharpness metric (95.4% sensitivity, 91.0% accuracy, 94.0% AUC). The highest classification performance is achieved by the final combination of Haralick features, sharpness metric and clustering (75.4% sensitivity, 91.7% accuracy, 94.8% AUC). The increase of the AUC of the Haralick features according to the growing number of additional feature types can be visualized by consulting the ROC curves of all feature subsets (figure 3a). The Haralick features show the highest p-value of 0.368 among the isolated feature groups compared to the final combination of all feature groups. The p-value increases in combination with clustering or sharpness features up to 0.368.

Comparing the data to ISC on our data set the final feature combination of the proposed method reaches a higher sensitivity than ISC (96.9% vs. 78.5%), a higher accuracy (91.7% vs. 86.7%) and a higher AUC (95.3% vs. 87.2%) (figure 3b). The isolated feature groups' AUC and the AUC of the combination of the sharpness metric and clustering do not significantly differ from ISC as their p-values lie above 0.05. The sharpness metric is an exception due to its worse curve and bad performance. The Haralick features having the strongest impact on the proposed method show a shrinking p-value by adding the other feature groups. Adding the sharpness metric to the Haralick features yields to a statistically different p-value of 0.037. The significantly lowest p-value of 0.019 is gained by using all three feature groups together.

Table 1 Evaluation results for classifying *bad* images. All possible feature combinations of the proposed method have been applied and compared to the state of the art approach *Image Structure Clustering* (ISC). Sensitivity at a specificity of 80.0%, accuracy at a classification threshold of 50.0%, area under the ROC curve (AUC) and p-value compared to ISC and compared to the combination of all features of the proposed method (HCS) are listed. The classification performance of the Haralick features (highest performance of isolated feature groups) grows with the number of combined features. In the same way shrinks the p-value relative to ISC and grows relative to the final combination of the proposed method's features (HCS). The proposed method's final feature combination (HCS) outperforms ISC gaining a higher sensitivity, accuracy and AUC.

Features	Sensitivity (%) (Specificity of 80%)	Accuracy (%) (Threshold of 50%)	AUC (%)	p-value (ISC)	p-value (HCS)
ISC (state of the art)	78.5	86.7	87.2	-	0.019
Sharpness	46.2	79.4	60.9	0.000	0.000
Clustering	89.2	86.3	90.7	0.376	0.138
Haralick	90.7	89.7	92.7	0.141	0.368
Clustering+Sharpness	87.7	86.4	89.3	0.597	0.067
Haralick+Clustering	93.9	90.4	94.0	0.058	0.641
Haralick+Sharpness	95.4	91.0	94.8	0.031	0.843
Haralick+Clustering+Sharpness	**96.9**	**91.7**	**95.3**	**0.019**	**-**

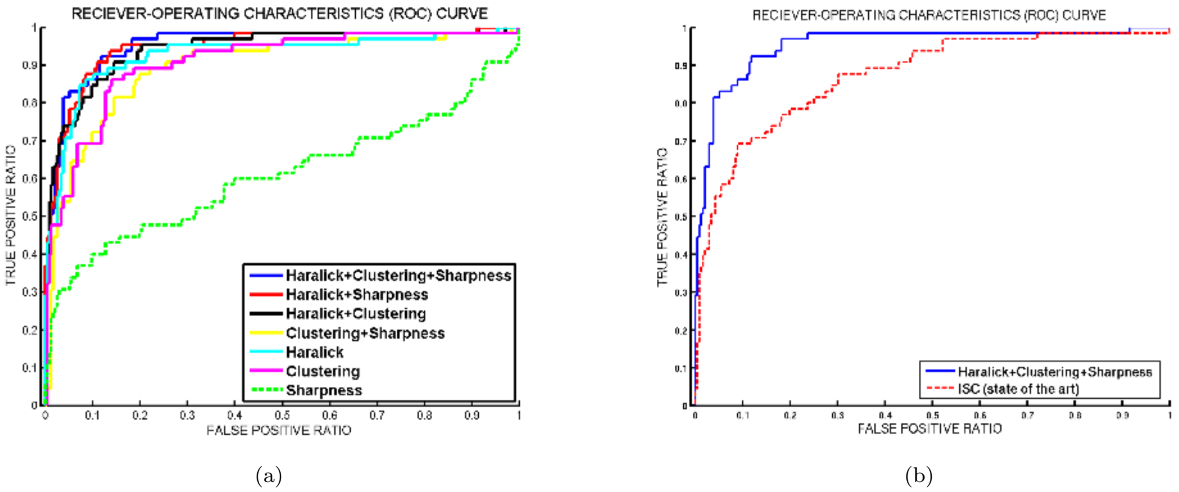


Fig. 3 Plots of ROC curves (finding *bad* images) for comparing the feature subsets of the proposed method to each other (a) and for comparing the proposed method with the state of the art *Image Structure Clustering* (ISC) (b). (a): For the isolated feature groups the Haralick features show the largest area under the ROC curve (AUC) and an increasing AUC in combination with growing types of information. (b): The proposed method shows a larger AUC compared to ISC indicating a higher classification performance.

We calculated Fleiss' κ by using the human observers' classification results and adding the automated methods' results as a fourth observer's results. The initial inter-observer-correlation ($\kappa = 0.58$) is increased for assuming the proposed method to be a fourth observer ($\kappa = 0.60$) and decreased for assuming ISC to be a fourth observer ($\kappa = -0.26$).

We measured the computation time per image for each feature group. The average computation time is 0.8 seconds for the sharpness metrics, 2.2 seconds for the clustering-features and 2.4 seconds for the Haralick features on an Intel Core 2 Duo Quad Q9550 system with 2.4 GHz and 3 GB RAM. No parallel processing was applied which results in a total computation time of 5.4 seconds. We compared this runtime to our implementation of the ISC-Algorithm as described in section 3.1.1 which takes 12.5 seconds in average running on the same machine without parallel processing. For both methods we evaluated only the computation time for generating the features. The classification step was omitted, since it has a comparable runtime for both approaches processing the same feature vector dimension.

4 Discussion and Conclusions

The proposed criteria inspired by diagnosis procedures based on the advise of an eye expert help to describe image quality objectively in the application of ophthalmology. The criteria are based on the recognizability and dissimilarity of certain structures in the eye background as well as on illumination homogeneity and sharpness. Our method models these criteria by the use of clustering, sharpness and Haralick features. The clustering detects certain components and measures how they can be recognized and differentiated. The sharpness metric calculates the edges' strength and evaluates how clearly the components are separated from each other. The Haralick features entropy, energy, and contrast are indicators for the generic image quality criteria, sharpness, homogeneity and contrast. The Haralick features have the strongest impact on the classification results with an AUC of 92.7%. Their performance is improved by combining all three feature groups. This is emphasized by the p-values related to the final combination of all features of proposed method. The Haralick features reach the highest p-value of 0.368 among the isolated feature groups. The p-value is increased by incorporating the clustering or the sharpness features. It shows higher values than the p-value of the combination of clustering and sharpness without Haralick features. All feature groups except sharpness and all combinations are not statistically different from the final combination of all features of the proposed method. With a resulting AUC of 95.3% compared to 87.2% the proposed method outperforms the state of the art ISC and shows a significant statistical difference (p-value = 0.019). Since ISC focuses on local gradient information for the structural clustering, it is not able to implement our required criteria on our data set. We have to state that in [12] ISC was designed and evaluated on a different gold standard using different quality criteria and a data set consisting of a wider FOV. The parameter set for the classifier found by our grid search strategy is comparable to the parameters in [12] (same penalty factor). Its clustering shows more detailed results especially in marking vessels (figure 2).

We presented a method that automatically quantifies the quality of retinal fundus images and produces reliable and stable results. The subjective understanding of quality could be defined objectively by introducing quality criteria. Nevertheless it still remains a hard task to classify an image at the class border, even for human evaluators and experts. Our method reaches an accuracy of 91.7% (96.9% sensitivity at a specificity of 80.0% for finding *bad* images) and an AUC of 95.3% by modeling our quality criteria. It outpeforms the state of the art approach ISC (50.8% sensitivity, 96.6% specificity, 86.7% accuracy, 87.2% AUC) on our data set. The evaluation is based on a manually pre-classified evaluation set of 301 images using cross-validation. We could show that the combination of local and global image statistics produces reliable and robust results in determining the image quality of retinal fundus photos and increases the sensitivity. This is important to identify *bad* images and to cause a reacquisition on the fly. Since both methods are designed for assessing the overall image quality, small local distortions like flash artifacts not affecting the quality significantly will not always be covered. The average feature computation time of 5.4 seconds per image for non-parallel processing is faster than comparable approaches. Our automated classification reaches the same correlation level as among the human observers. Thus, the proposed method is closer to a human decision than other approaches. We assume that our features model the human perception by implementing our criteria. The method evaluates the recognizability for diagnosis relevant structures like it is perceived by experts. It also judges generic image quality with similar criteria used by the human perception. This seems not to be the case for ISC.

As a conclusion we can state that we developed a method to automatically assess retinal fundus image quality. By introducing relevant criteria the objectivity of individual human perception for quality could be improved. But in particular at the class border the discrimination of *good* and *bad* images remains a crucial task. Our method underlies the same restrictions as it is limited by the the human graded gold standard. Nevertheless, we can substitute an human quality evaluation by the fast objective measurement presented here to ensure a sufficient image quality level in broad screening applications.

Acknowledgements The authors gratefully acknowledge funding of the Erlangen Graduate School in Advanced Optical Technologies (SAOT) by the German Research Foundation (DFG) in the framework of the German excellence initiative. The work of J. Meier was supported by a scholarship from the International Max Planck Research School on Optics and Imaging, Erlangen.

References

1. M. D. Abràmoff, M. Niemeijer, M. S. Suttorp-Schulten, M. A. Viergever, Stephen R. Russell and Bram van Ginneken, Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes, Diabetes Care, 31 (2), 193–198 (2008)
2. R. Bock, J. Meier, L. G. Nyúl, J. Hornegger and G. Michelson, Glaucoma risk index: Automated glaucoma detection from color fundus images, Medical Image Analysis, 14 (3), 471-481 (2010)
3. C. Sinthanayothin, J. F. Boyce, T. H. Williamson, H. L. Cook, E. Mensah, S. Lal and D. Usher, Automated detection of diabetic retinopathy on digital fundus images. Diabetic Medicine, 19 (2), 105-112 (2002)
4. M. D. Abràmoff and M. Suttorp-Schulten, Web-based screening for diabetic retinopathy in a primary care population: the

eye check project. Telemedicine and e-Health, 11 (6), 668-674 (2005)

5. A. M. Eskicioglu and P. S. Fisher, Image quality measures and their performance, IEEE Transactions on Communications, 3 (12), 2959–2965 (1995)
6. İ. Avcıbaş, B. Sankur and K. Sayood, Statistical evaluation of image quality measures, Journal of Electronic Imaging, 11 (2), 206–223 (2002)
7. Z. Wang, A. C. Bovik and L. Lu, Why is image quality assessment so difficult?, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 4, 3313–3316 (2002)
8. 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, Investigative Ophthalmology and Visual Science, 47 (3), 1120–1125 (2006)
9. L. Giancardo, M. D. Abràmoff, E. Chaum, T. P. Karnowski, F. Meriaudeau and K. W. Tobin Jr, Elliptical local vessel density: a fast and robust quality metric for retinal images, Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, 3534–3537 (2008)
10. M. Lalonde, L. Gagnony and M.-C. Boucher, Automatic visual quality assessment in optical fundus images, Proceedings of Vision Interface (VI 2001), 259–264 (2001)
11. S. C. Lee and Y. Wang, Automatic retinal image quality assessment and enhancement, Proceedings of SPIE, 3661, 1581–1590 (1999)
12. M. Niemeijer, M. D. Abràmoff and B. van Ginneken, Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening, Medical Image Analysis, 10 (6), 888–898 (2006)
13. R. M. Haralick, K. Shanmugam and I. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man and Cybernetics, 3 (6), 610–621 (1973)
14. C.-C. Chang and C.-J. Lin, “LIBSVM”: a library for support vector machines, http://www.csie.ntu.edu.tw/∼cjlin/libsvm (2001)