Automatic No-Reference Quality Assessment for Retinal Fundus Images Using Vessel Segmentation

Thomas Köhler[1,2], Attila Budai[1,2], Martin F. Kraus[1,2], Jan Odstrčilik[4,5],
Georg Michelson[2,3], Joachim Hornegger[1,2]
[1]Pattern Recognition Lab, University of Erlangen-Nuremberg, Erlangen, Germany
[2]Erlangen Graduate School in Advanced Optical Technologies (SAOT), Erlangen, Germany
[3]Department of Ophthalmology, University of Erlangen-Nuremberg, Erlangen, Germany
[4]Department of Biomedical Engineering, Brno University of Technology, Brno, Czech Republic
[5]St. Anne's University Hospital - International Clinical Research Center (ICRC), Brno, Czech Republic
thomas.koehler@fau.de

Abstract

Fundus imaging is the most commonly used modality to collect information about the human eye background. Objective and quantitative assessment of quality for the acquired images is essential for manual, computer-aided and fully automatic diagnosis. In this paper, we present a no-reference quality metric to quantify image noise and blur and its application to fundus image quality assessment. The proposed metric takes the vessel tree visible on the retina as guidance to determine an image quality score. In our experiments, the performance of this approach is demonstrated by correlation analysis with the established full-reference metrics peak-signal-to-noise ratio (PSNR) and structural similarity (SSIM). We found a Spearman rank correlation for PSNR and SSIM of 0.89 and 0.91. For real data, our metric correlates reasonable to a human observer, indicating high agreement to human visual perception.

1 Introduction

Fundus imaging is the most commonly used modality to collect information about the human eye background for the diagnosis of various retinal diseases such as glaucoma or diabetic retinopathy. Retinal image analysis is an active field of research providing image processing and machine learning methods either for computer-assisted or fully automatic diagnoses [1]. However, for the success of these methods high-quality image data is essential. Images of poor quality must be detected by an operator and the acquisition must be repeated, which is a highly subjective decision and a time-consuming task. Furthermore, image processing techniques ranging from autofocus [6] to image deconvolution [5] recently established in ophthalmic imaging deal with images of different quality that must be quantitatively assessed to detect most the most reliable image (e. g. for autofocus) or to evaluate image improvement (e. g. in deconvolution).

No-reference image quality assessment deals with the problem to provide a quantitative score of image quality in the absence of a gold standard. In literature, there exist two groups of methods for solving this task: (i) classification-based approaches and (ii) quality metrics for image content. In methods falling in category (i), image quality is predicted by assigning an image to one class out of a discrete set of quality classes using supervised learning strategies. This is achieved using feature extraction and classification based on a gold standard provided by experts [8, 9]. Even if such methods are attractive for identifying good images for diagnosis purposes, the application is limited to problems where a discrete assessment is sufficient. (ii) Quality metrics are scores for general quality features like image noise or sharpness to provide a continuous measure in an unsupervised manner, which is the focus in this paper. For the prediction of the relative amount of sharpness in natural images Narvekar and Karam [7] proposed the cumulative probability of blur detection (CPBD). A quality metric to estimate image noise and blur simultaneously is Renyi entropy [3] adopted to fundus imaging by Marrugo et al. [4]. A major limitation of these methods is that a uniform quality across the whole image is assumed which is not always valid in case of fundus images. This is caused by the curvature of the retina or diseases which introduce local blur. Zhu et al. [12] proposed a novel metric to quantify noise and blur which does not require uniform disturbances.

In this work, we focus on simultaneous quantification of image blur and noise. Here, we adopt the approach originally introduced by Zhu et al. [12] for automatic qual-

ity assessment of retinal fundus images. In section 2 we present the state-of-the-art approach applicable to each image modality. Our specialized approach for fundus images is introduced in section 3 and takes the vessel tree as guidance to determine an objective and continuous quality score. The performance is demonstrated by analyzing the correlation between no-reference assessment and established full-reference metrics in section 4. This method may be used either as stand-alone metric to detect blurred and noisy images or as feature in a classification-based approach.

2 Background

Let $\boldsymbol{I}$ be a grayscale image of size $M \times N$. We decompose $\boldsymbol{I}$ in a set of distinct patches, whereas each patch $\boldsymbol{P}$ is of size $n \times n$. The local gradient matrix $\boldsymbol{G}$ of size $n^2 \times 2$ for $\boldsymbol{P}$ is given by:

$$\boldsymbol{G} = \begin{pmatrix} P_x(1,1) & P_y(1,1) \\ \vdots & \vdots \\ P_x(n,n) & P_y(n,n) \end{pmatrix}, \tag{1}$$

where $P_x(x_i, y_i)$ and $P_y(x_i, y_i)$ denotes the derivative of $\boldsymbol{P}$ at pixel (x_i, y_i) in x- and y-direction, respectively. The singular value decomposition (SVD) of $\boldsymbol{G}$ is given by:

$$\boldsymbol{G} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \tag{2}$$

$$= \boldsymbol{U} \begin{pmatrix} s_1 & 0 \\ 0 & s_2 \end{pmatrix} \boldsymbol{V}^T, \tag{3}$$

for orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ and the singular values s_1, s_2. It is shown in [12] that a local quality metric to quantify image noise and blur in an anisotropic patch $\boldsymbol{P}$ is given by:

$$q(\boldsymbol{P}) = s_1 \cdot R, \tag{4}$$

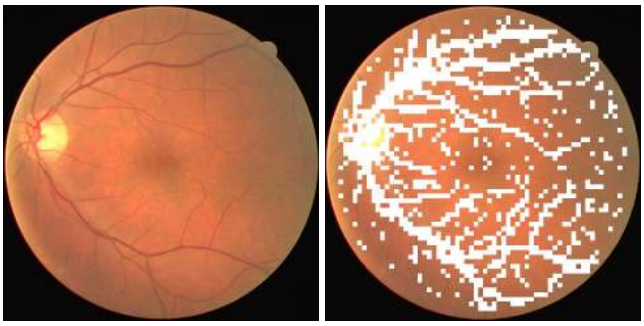
where R denotes the coherence:

$$R = \frac{s_1 - s_2}{s_1 + s_2}. \tag{5}$$

Larger values for $q(\boldsymbol{P})$ defined in (4) indicate higher image quality in terms of blur and noise. It is important to note, that $q(\boldsymbol{P})$ is only a valid quality metric in an anisotropic patch with dominant gradient direction, whereas in isotropic patches the score is not meaningful. In order to get a global estimate for noise and blur, $q(\boldsymbol{P})$ is summed up over all anisotropic patches and normalized according to:

$$Q = \frac{1}{MN} \sum_{i,j:\mathcal{P}(i,j)=1} q(\boldsymbol{P}_{ij}) \tag{6}$$

where $\mathcal{P}(i,j)$ denotes the patch map for image $\boldsymbol{I}$ such that $\mathcal{P}(i,j) = 1$ if $\boldsymbol{P}_{ij}$ is anisotropic. These patches may be detected automatically employing statistical tests for the coherence R (see Fig. 1). Patches having significant coherences $R > \tau_R$ are assumed to be anisotropic for a fixed threshold τ_R calculated by:



(a) (b)

Figure 1: A color fundus image (a) and the detected anisotropic patches of size 8×8 pixels (b).

$$\tau_R = \sqrt{\frac{1 - \alpha^{\frac{1}{n^2-1}}}{1 + \alpha^{\frac{1}{n^2-1}}}}, \tag{7}$$

where α is the significance level for testing if a given patch is anisotropic. For the patch size we set $n = 8$ and for the significance level $\alpha = 0.001$ as suggested in [12].

3 Proposed Method

In this section, we describe, how the no-reference quality metric Q is applied to color fundus images. Next, we propose an extended approach which takes blood vessels visible in fundus images as guidance to determine a spatially weighted quality score.

3.1 Color Image Quality Assessment

The quality metric Q given by Eq. (6) is defined for grayscale images only. However, in fundus imaging a quality score for color images is required. Here, contrast and saturation of the color channels in RGB space are different. Usually the blue channel has poor contrast between background and anatomical structures, whereas the red channel is often overexposed. However, we assume uniform quality of the channels with respect to noise and sharpness. Therefore we propose to extract the green color channel for quality assessment, since it shows the best contrast and provides maximal structure for subsequent quality assessment.

3.2 Vessel-Based Quality Assessment

One limitation of the quality metric Q is the automatic detection of anisotropic patches. Using thresholding procedures may lead to false selections, especially in the case

of noisy or highly blurred images. On the other hand, fundus images consist of relatively few texture and structure compared to natural images. Thus, the number of possible candidates for anisotropic patches is low. To overcome this problem, we propose to use the vessel tree as guidance, since we expect that blood vessel boundaries are good candidates for true anisotropic patches.

3.2.1 Vesselness Measure

We detect blood vessels in an image $\boldsymbol{I}$ as follows. First, the green color channel $\boldsymbol{I}_g$ is extracted from the color image $\boldsymbol{I}$ due to the good contrast between vessels and background compared to the other channels. For each pixel in $\boldsymbol{I}_g$ the local Hessian matrix is calculated by:

$$\boldsymbol{H} = \begin{pmatrix} \frac{\partial d^2 \boldsymbol{I}_g}{\partial x^2} & \frac{\partial d^2 \boldsymbol{I}_g}{\partial x \partial y} \\ \frac{\partial d^2 \boldsymbol{I}_g}{\partial x \partial y} & \frac{\partial d^2 \boldsymbol{I}_g}{\partial y^2} \end{pmatrix}. \tag{8}$$

For detection of blood vessels, we employ the *vesselness* measure proposed by Frangi et al. [2] according to:

$$V = \exp\left(-\frac{\lambda_1^2}{\lambda_2^2}\right)\left(1 - \exp\left(-(\lambda_1^2 + \lambda_2^2)\right)\right) \tag{9}$$

for the eigenvalues λ_1 and λ_2 of $\boldsymbol{H}$ where $\lambda_2 \geq \lambda_1$. Here, V represents a probability measure where large values indicate high probability for pixels to be located on a vessel (see Fig. 2). Since we are mainly interested in thick vessels and in order to decrease noise in the vesselness map, we neglect pixels having small vesselness and set $V = 0$ for $V < V_0$ below a fixed threshold V_0. We set V_0 adaptively to the 80^{th} percentile of all non-zero vesselness measurements.

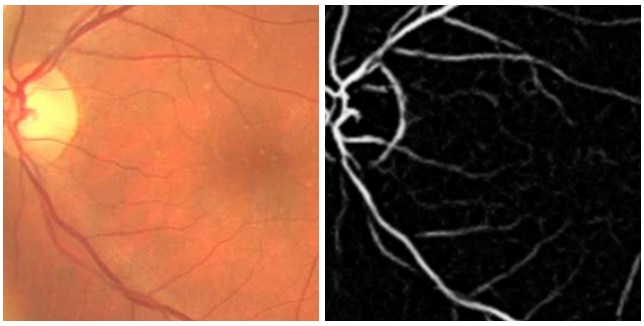
In its original version, vesselness is computed pixel-wise according to (9) using different window sizes to determine the Hessian. Then, the size achieving the largest vesselness is used for vessel detection. In this paper, we make use of a multi-scale approach and determine the vesselness for a fixed window of size 3×3 pixel but at downsampled versions of the original image. This method speeds up the computation of V for vessel detection.

3.2.2 Spatially Weighted Quality Metric

We utilize the vesselness as spatially adaptive confidence weight for quality assessment of retinal fundus images. Here, the basic idea is that anisotropic patches located on blood vessel boundaries are more reliable for the overall blur and noise estimate. Our vessel-based quality metric is defined as:

$$Q_v = \sum_{i,j:\mathcal{P}(i,j)=1} \tilde{\Sigma}_{ij} \cdot q(\boldsymbol{P}_{ij}), \tag{10}$$

where $\tilde{\Sigma}_{ij}$ denotes the normalized local variance of the vesselness measure in patch $\boldsymbol{P}_{ij}$. Here, $\tilde{\Sigma}_{ij}$ is determined by computing the variance of the vesselness V in $\boldsymbol{P}_{ij}$, where normalization is done using the overall patch number such that $\sum_{i,j} \tilde{\Sigma}_{ij} = 1$. Thus, patches $\boldsymbol{P}_{ij}$ located on a blood vessel boundary indicated by large $\tilde{\Sigma}_{ij}$ have higher impact to the overall estimate for image noise and blur. We demonstrate in our experiments that this quality score is more reliable than thresholding for patch detection and the use of uniform weights for all patches.



(a) (b)

Figure 2: An example color image (a) and calculated vesselness measure for blood vessel detection (b).

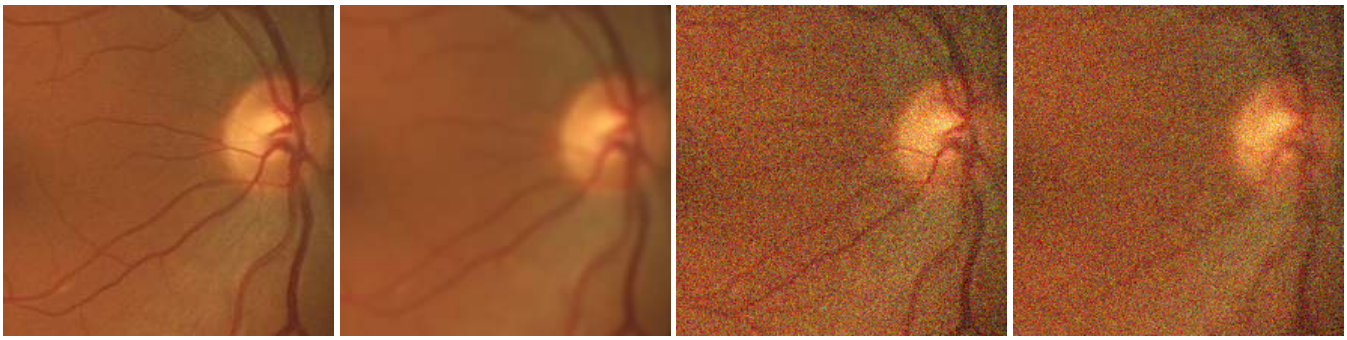
4 Experiments and Results

We evaluated the ability of the proposed quality metric to quantify sharpness and noise in retinal fundus images. First, our Q_v defined in (10) is compared to the original score Q defined in (6) by analyzing the agreement with full-reference metrics based on synthetic images. We also evaluated our approach for real image data. Supplementary material for our experiments is available on our web page[1].

4.1 Correlation to Full-Reference Metrics

For quantitative evaluation, we used 40 images out of the DRIVE database [10]. From each original image we generated synthetic images in two steps: (i) We induced blur using a Gaussian filter with fixed size of 7×7 and varying standard deviation σ_b. (ii) From each blurred image, a noisy image was generated by adding zero-mean Gaussian noise of varying standard deviation σ_n (see Fig. 3). Having an original image $\boldsymbol{I}$ and a disturbed image $\tilde{\boldsymbol{I}}$, we employ the established full-reference quality metrics peak-signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [11] to quantify the degradation of $\tilde{\boldsymbol{I}}$. For evaluation of the reliability of the proposed metric, we calculated Spearman's rank correlation ρ between the full-reference metrics and the no-reference quality scores.

[1] http://www5.cs.fau.de/en/our-team/koehler-thomas



(a) (b) (c) (d)

Figure 3: Region of interest of an image used as ground truth (a), generated blurred image ($\sigma_b = 3.0$) (b), generated noisy image ($\sigma_n = 10^{-2}$) (c) and generated blurred and noisy image (d).

4.1.1 Assessment of Blur

In our first experiment, we varied the amount of Gaussian blur from $\sigma_b = 0.5$ to $\sigma_b = 3.0$. For each blur level, 20 different noise levels ranging from $\sigma_n = 10^{-4}$ to $\sigma_n = 10^{-2}$ were simulated. Both parameters were increased logarithmically to achieve a uniform sampling of the corresponding PSNR and SSIM measures. Spearman's ρ was calculated for all blur levels between each no-reference quality score (Q and Q_v) and PSNR as well as SSIM. Mean and standard deviation of ρ averaged over 40 images are plotted in Fig. 4. If σ_b becomes large, ρ is decreased on average and has a higher standard deviation. Even for large σ_b we achieve correlations higher than 0.8 between the full-reference metrics and the no-reference metrics. Please note, that correlations for PSNR and SSIM were equal in our experiment, since both scores were perfectly correlated.

4.1.2 Assessment of Image Noise

We repeated our first experiment and calculated Spearman's ρ for different noise standard deviations σ_n. Mean and standard deviation for ρ averaged over 40 images are plotted in Fig. 5. Here, Spearman's ρ decreased and had a higher standard deviation if the noise level σ_n was increased. However, for moderate noise levels ($\sigma_n \approx 3 \cdot 10^{-3}$) we still achieve correlations higher than 0.6 for Q and Q_v.

4.1.3 Overall Correlation

We also analyzed Spearman's ρ over a whole experiment, where image noise and blur were varied simultaneously for 40 images as well as 20 noise and blur levels, respectively. A comparison between Q and Q_v is shown in Tab. 1. Here, we achieve a Spearman correlation of above 0.8 for both approaches with respect to the full-reference metrics PSNR and SSIM.

Table 1: Spearman's ρ for simultaneously varying noise and blur for metric Q and our proposed metric Q_v.

Full-ref. metric	$\rho(Q)$	$\rho(Q_v)$
PSNR	0.8227	0.8920
SSIM	0.8412	0.9076

4.2 Real Images

We captured 18 image pairs of the same eye from 18 human subjects using a Canon CR-1 fundus camera with a field of view of 45°. For each pair, the first image suffers from decreased sharpness and thus the examination had to be repeated. Both images share approximately the same field of view, whereas small shifts were caused by eye movements between the acquisitions (see Fig. 6). The proposed metric Q_v was compared to the original Q metric [12] as well as to the CPBD metric [7] and the anisotropy measure [3]. All metrics were applied to the field of view whereas the background regions were masked out for quality assessment. For normalization, we used the $m = 10^4$ most significant anisotropic patches to determine Q and Q_v for each image, to neglect the effect of the patch number to the global quality score.

We considered quality classification implemented as thresholding of the estimated quality score. ROC curves for classification based on the different metrics are shown in Fig.7. For Q_v we obtained an area under the ROC curve of 88.3% (CPBD: 50.9%, Anisotropy: 75.3%, Q: 79.6%).

The metric Q_v was also compared pair-wise between a good acquisition and the corresponding image of poor quality. Here, the ranking obtained by Q_v agrees to a human observer for 16 out of 18 image pairs resulting in an agreement of 88.9% (CPBD: 55.6%, Anisotropy: 94.4%, Q: 83.3%).

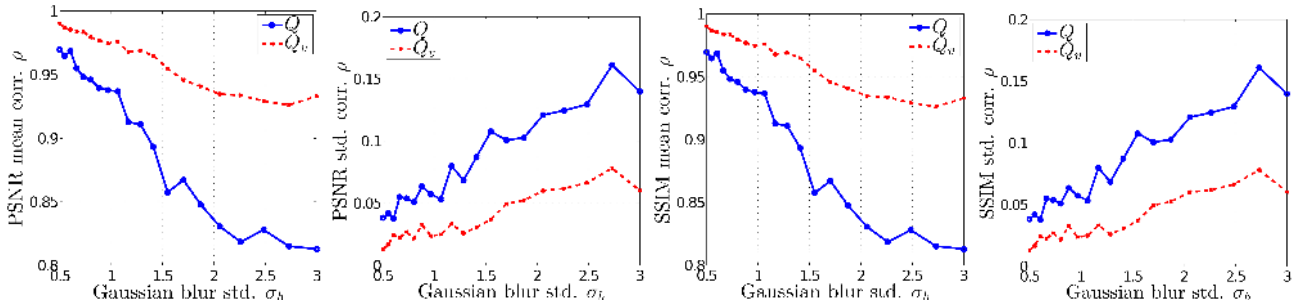


Figure 4: Mean and standard deviation of Spearman's ρ between no-reference quality assessment and PSNR ((a) and (b)) as well as SSIM ((c) and (d)) for 40 test images versus varying amount of artificial blur.

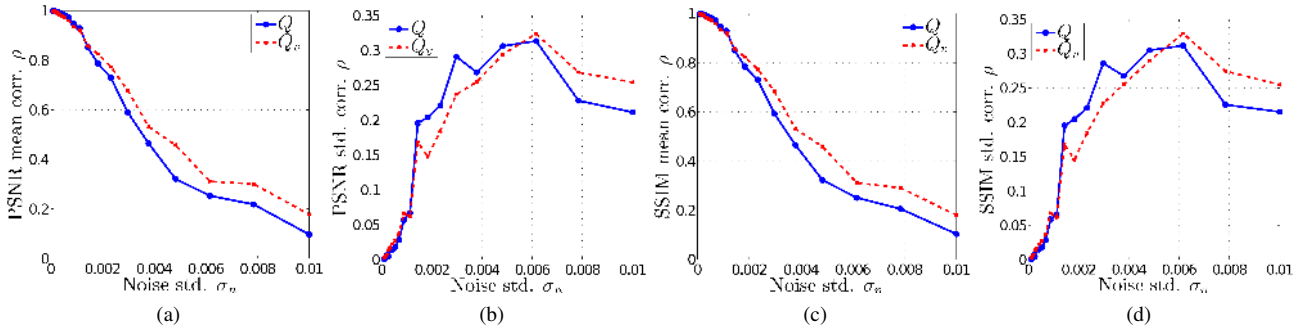


Figure 5: Mean and standard deviation of Spearman's ρ between no-reference quality assessment and PSNR ((a) and (b)) as well as SSIM ((c) and (d)) for 40 test images versus varying amount of additive Gaussian noise.

4.3 Discussion

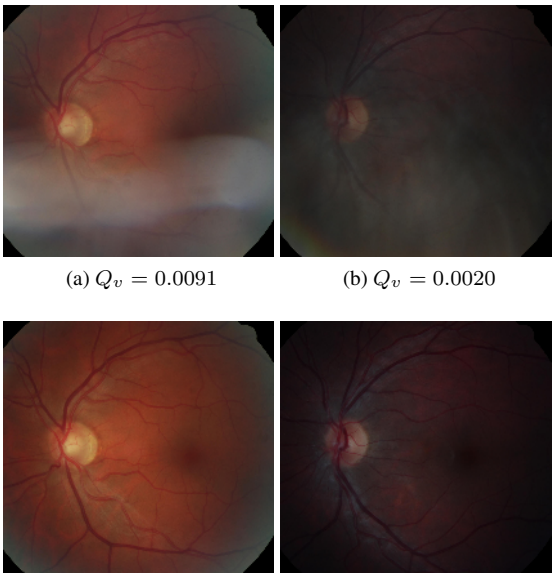
As shown in Fig. 4 for each blur level, our metric Q_v outperforms the original approach with respect to mean and standard deviation of Spearman's ρ. This is especially noticeable for high amounts of blur. In contrast to this result, in the case of varying noise levels shown in Fig. 5, ρ is lower for both Q and Q_v. However, the mean correlation for Q_v is still improved. Spearman's ρ for simultaneously varying noise and blur summarized in Tab.1 indicates higher correlations of Q_v to the full-reference metrics with significance level 0.05. Thus, Q_v has a higher agreement with both full-reference metrics over a wide range of noise and blur.

In our experiments using real images, Q_v agrees reasonable with visual inspection of the camera's operator. This is also the case for non-uniform degradations such as spatially varying blur (see Fig. 6a). In terms of quality classification, our approach outperforms state-of-the-art methods indicated by an improved area under the ROC curve. Please note, that our method measures blur and noise whereas related aspects such as illumination homogeneity is not explicitly taken into account. However, Q_v may be combined with various features to assess different quality criteria.

5 Conclusion

In this paper, we presented an improved no-reference image quality metric Q_v to quantify the amount of noise and blur in retinal fundus images. For reliable quality estimation, we employ the vessel tree detected by the well known vesselness measure as guidance to determine a global quality score from local estimates in anisotropic patches. The proposed metric shows high agreement with the established full-reference metrics PSNR and SSIM indicated by a Spearman rank correlation of 0.89 and 0.91, respectively. Thus, Q_v is able to replace full-reference metrics for quality assessment in the absence of a gold standard. For real data, our metric agrees reasonable to visual inspection of a human operator in terms of image sharpness.

In our future work, we will study the adaption of the proposed method to applications where image sharpness has to be continuously assessed, such as camera auto-focusing. As another application we focus on the integration of Q_v as feature into a classification-based quality rating in combination with different quality features. Our experiments using real data indicates that this may be feasible and an extensive evaluation on large image databases is ongoing research.



(a) $Q_v = 0.0091$ (b) $Q_v = 0.0020$

(c) $Q_v = 0.0122$ (d) $Q_v = 0.0099$

Figure 6: Fundus images and corresponding scores Q_v: If the quality of the first acquisition was too low (first row), the examination was repeated (second row). Images of poor quality suffer either from local loss of sharpness (a) or are globally degraded (b).

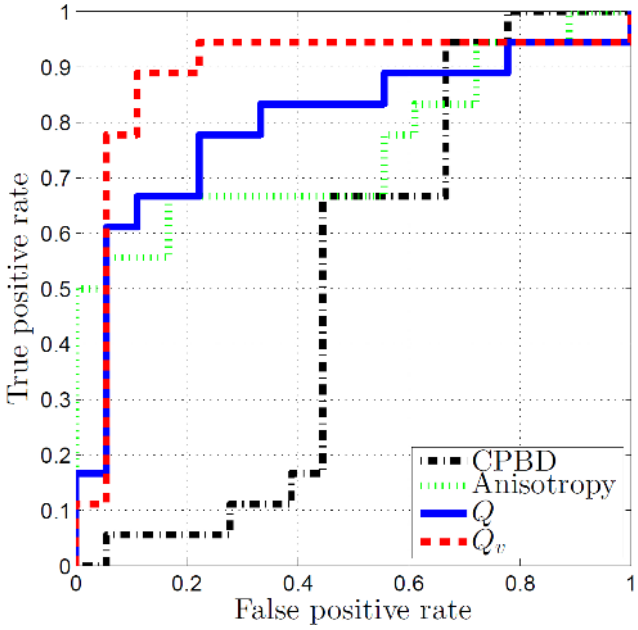


Figure 7: ROC curve for quality classification based on different metrics.

Acknowledgment The authors gratefully acknowledge funding of the Erlangen Graduate School in Advanced Optical Technologies (SAOT) by the German National Science Foundation (DFG) in the framework of the excellence initiative. This project is supported by the German Federal Ministry of Education and Research, project grant No. 01EX1011D, the European Regional Development Fund - Project FNUSA-ICRC (No. CZ.1.05/1.1.00/02.0123) and by the Czech-German project no. 7AMB12DE002 under Ministry of Education, Youth and Sports.

References

[1] M. D. Abramoff, M. K. Garvin, and M. Sonka. Retinal Imaging and Image Analysis. *IEEE Reviews in Biomedical Engineering*, 3:169–208, 2010.

[2] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever. Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 1998*, volume 1496 of *Lecture Notes in Computer Science*, pages 130–137. Springer Berlin Heidelberg, 1998.

[3] S. Gabarda and G. Cristóbal. Blind image quality assessment through anisotropy. *Journal of the Optical Society of America A*, 24(12):B42–B51, 2007.

[4] A. G. Marrugo, M. S. Millán, G. Cristóbal, S. Gabarda, and H. C. Abril. No-reference quality metrics for eye fundus imaging. In *Proceedings of the 14th international conference on Computer analysis of images and patterns - Part I*, pages 486–493, Seville, Spain, 2011. Springer-Verlag.

[5] A. G. Marrugo, M. Sorel, F. Sroubek, and M. S. Millán. Retinal image restoration by means of blind deconvolution. *Journal of Biomedical Optics*, 16(11):116016, 2011.

[6] M. Moscaritolo, H. Jampel, F. Knezevich, and R. Zeimer. An image based auto-focusing algorithm for digital fundus photography. *IEEE Transactions on Medical Imaging*, 28(11):1703–1707, 2009.

[7] N. Narvekar and L. Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *Image Processing, IEEE Transactions on*, 20(9):2678–2683, 2011.

[8] 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.

[9] J. Paulus, J. Meier, R. Bock, J. Hornegger, and G. Michelson. Automated quality assessment of retinal fundus photos. *International Journal of Computer Assisted Radiology and Surgery*, 5(6):557–564, 2010.

[10] J. 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 Transactions on Medical Imaging*, 23:501–509, 2005.

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[12] X. Z. X. Zhu and P. Milanfar. Automatic Parameter Selection for Denoising Algorithms Using a No-Reference Measure of Image Content. *IEEE Transactions on Image Processing*, 19(12):3116–3132, 2010.