Available online at www.sciencedirect.com
ScienceDirect
Medical Image Analysis 10 (2006) 888–898
MEDICAL IMAGE ANALYSIS
www.elsevier.com/locate/media

Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening

Meindert Niemeijer [a,*,1], Michael D. Abràmoff [b,2], Bram van Ginneken [a]

[a] University Medical Center Utrecht, Image Sciences Institute, Q0S.459, Heidelberglaan 100, 3584CX, Utrecht, The Netherlands
[b] Department of Ophthalmology and Visual Sciences, University of Iowa Hospitals and Clinics, 200 Hawkins Drive, Iowa City, IA 52242, USA

Received 20 March 2006; received in revised form 20 July 2006; accepted 14 September 2006

Abstract

Reliable verification of image quality of retinal screening images is a prerequisite for the development of automatic screening systems for diabetic retinopathy.

A system is presented that can automatically determine whether the quality of a retinal screening image is sufficient for automatic analysis. The system is based on the assumption that an image of sufficient quality should contain particular image structures according to a certain pre-defined distribution. We cluster filterbank response vectors to obtain a compact representation of the image structures found within an image. Using this compact representation together with raw histograms of the R, G, and B color planes, a statistical classifier is trained to distinguish normal from low quality images. The presented system does not require any previous segmentation of the image in contrast with previous work.

The system was evaluated on a large, representative set of 1000 images obtained in a screening program. The proposed method, using different feature sets and classifiers, was compared with the ratings of a second human observer. The best system, based on a Support Vector Machine, has performance close to optimal with an area under the ROC curve of 0.9968.

Keywords: Image quality; Retina; Screening; Diabetic retinopathy; Image structure

1. Introduction

Diabetic Retinopathy (DR) is an ocular complication of diabetes. It is the most important cause of blindness in the working population of the European Union and the United States (Klonoff and Schwartz, 2000). It has been shown that early diagnosis and timely treatment can prevent vision loss in most cases. Yet, for a variety of reasons, less than 50% of diabetics are screened for the presence of DR. One important limiting factor is the required scale of a program to screen the entire population of diabetics. For example, in the US alone, 18 million people would have to have their eyes examined annually. Computer aided diagnosis technology could facilitate such a large program. Most of the images in a screening program do not contain abnormalities. For example, in the screening program that supplied the data used in this work less than 10% of all subjects showed signs of DR (Abràmoff and Suttorp-Schulten, 2005). Our research is focused on the development of a computerized DR pre-screening system. The system would make a selection of images that possibly contain DR and present only these cases to the ophthalmologist. This could reduce the overall workload of ophthalmologists significantly.

* Corresponding author. Tel.: +31 30 250 4635; fax: +31 30 251 3399.
E-mail address: meindert@isi.uu.nl (M. Niemeijer).
[1] Meindert Niemeijer was supported by the Dutch Ministry of Economic Affairs through IOP IBVA02016.
[2] Michael Abràmoff is supported by the National Eye Institute R01 EY017066, US Department of Agriculture Distance Learning Telemedicine program, Department of Defense STTR A06-T030, the Netherlands Organization for Health Related Research (ZonMW), Research to Prevent Blindness, NY, NY and the Wellmark Foundation.

1361-8415/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2006.09.006

In medical imaging in general, image quality is an important topic. However, the automatic detection of image quality is an avenue of research that has not received a lot of attention. Retinal images obtained in a screening program are acquired at different sites, using different cameras that are operated by qualified people who have varying levels of experience. This results in a large variation in image quality and a relatively high percentage of images with an insufficient quality. In the screening program that supplied the data used in this study, 12% of the images were marked as unreadable by the ophthalmologists (Abràmoff and Suttorp-Schulten, 2005). The quality of an image is deemed insufficient when it becomes difficult or impossible to make a meaningful clinical judgment regarding the presence or absence of signs of DR in the image (see Fig. 1 for example images). Performing computerized analysis on an image of insufficient quality will produce unreliable results. These low quality images should be examined by an ophthalmologist and reacquired if needed. For an automated DR screening system sensitivity is important, one wants to detect the first, subtle signs of the presence of DR. In many cases of low image quality the contrast is reduced. This could hide small abnormalities and cause the system to label the image as normal while abnormalities are present. An automatic image quality verification system is therefore a vital part of any automatic DR screening. The development and testing of such a quality verification system based upon a general method to describe image structures is the focus of this work.

Medical images are typically acquired according to a protocol. Therefore, one can make the a priori assumption that the type of structures and their relative ratios found in any image that has been acquired according to the same protocol are similar. Problems in the acquisition of the image resulting in low image quality generally result in a disturbance in the detected image structures. A general method we call Image Structure Clustering (ISC) is used to provide a compact representation of the structures found in an image. The technique determines the most important set of structures, present in a set of normal quality images, based on a clustering of the response vectors generated by a filterbank. Clustering of filterbank responses has been used for different applications in image processing. ISC is most similar to the texton-based approach to texture description (Malik et al., 2001). However, ISC utilizes a multiscale filterbank and thus allows detection of important image structures on multiple scales. As the technique is applied to medical images that all contain similar structures a low number of clusters suffices to describe the most important structures.

Major causes of low image quality in retinal screening images include loss of contrast due to movement of the patient or movement of the eye, non-uniform illumination of the retina due to insufficient pupil size, imaging of (part of) the eyelid due to blinking and opaque media preventing normal quality imaging of the retina. Example images are shown in Fig. 2.

Previously presented approaches for retinal image quality determination focussed on global image intensity histogram analysis (Lee and Wang, 1999) or analysis of the global edge histogram in combination with localized image intensity histograms (Lalonde et al., 2001). In both these approaches a small set of excellent quality images was used to construct a mean histogram. The difference between the mean histogram and a histogram of a given image then indicates the image quality. A shortcoming of these methods is that they use only a limited type of analysis and rely on one mean histogram for comparison that does not account for the natural variance encountered in retinal images.

Another approach is taken by Fleming et al. (2006) and Lowell et al. (2005), both present systems that are very

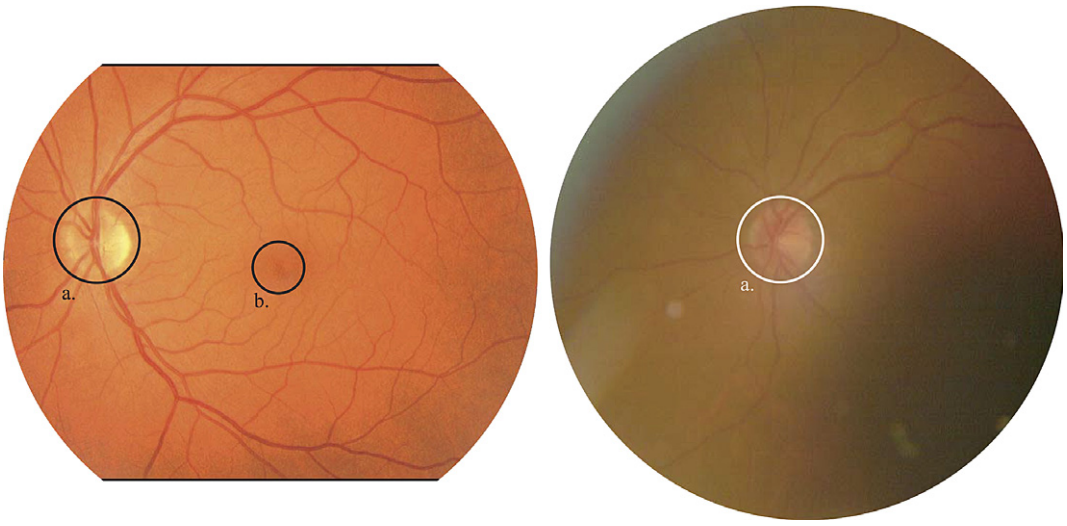

Fig. 1. Two example retinal images, of normal quality (left) and low quality (right). (a) Indicates the optic disc or "blind spot" where the optic nerve exits the eye. (b) Is the fovea where visual acuity occurs. In the right image the fovea is not visible.

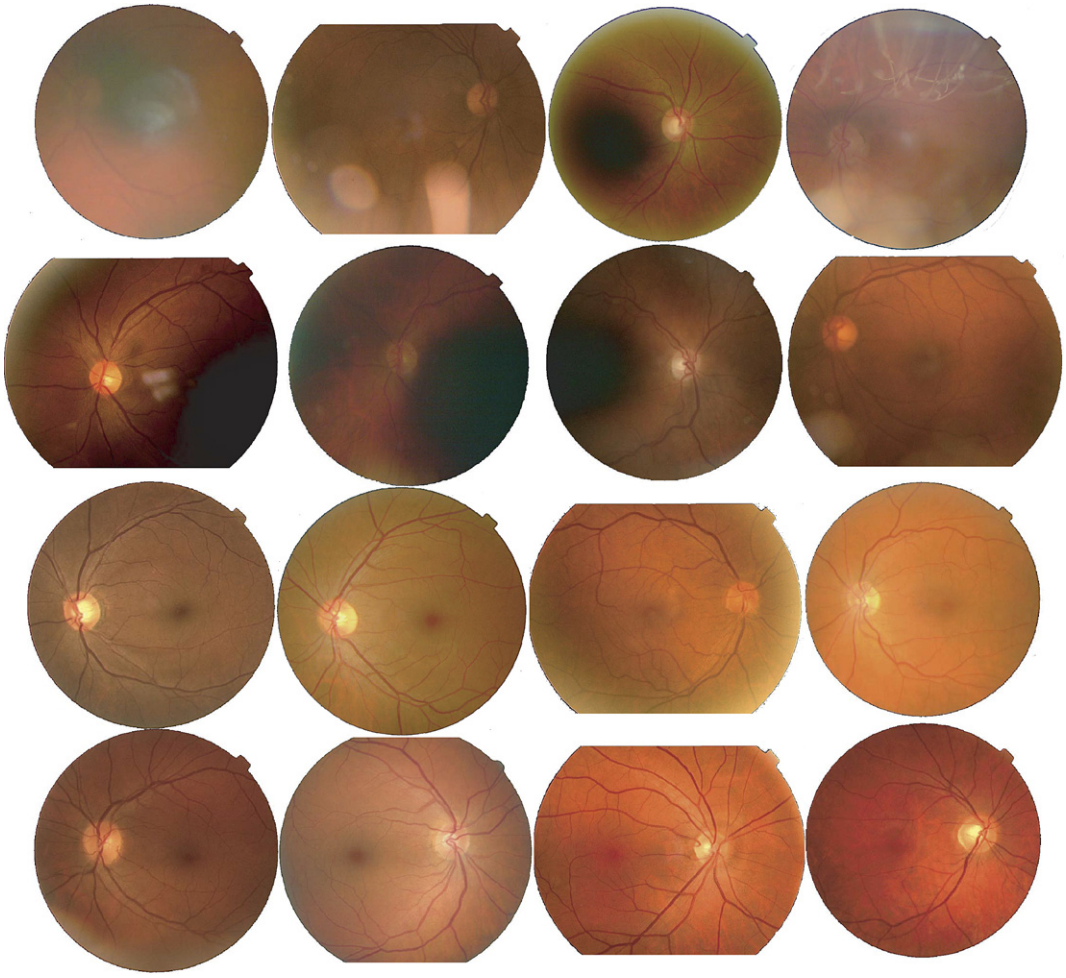

Fig. 2. The top two rows contain eight typical low quality images. The bottom two rows contain eight typical normal quality images.

specific for analysis of retinal images. Both systems classify images by assigning them to one of a number of quality classes. An analysis is made of the vasculature in a circular area around the macula. The presence of small vessels in this area is used as an indicator of image quality. The presented results are good, but the method requires a segmentation of the vasculature and other major anatomical structures to find the region of interest around the fovea. Detecting the failure of a segmentation in case of low image quality is not trivial. This paper presents a supervised method that learns which image structures are present in normal quality images and their relative ratios. For training and testing the system a large set of example images is used. The method does not require any specific analysis and is a general method that could be applied to other types of medical images as well.

The article is structured as follows. In Section 2 the image data and reference standard are described. The image structure clustering technique and its application to retinal images are presented in Section 3. In Section 4 the complete quality verification system is described. Section 5 presents the experimental results and the paper ends with a discussion and conclusion in Section 6.

2. Materials

A total of 2000 images were obtained from a DR screening program in the Netherlands. The demographic data of the screening program were consistent with an elderly type 2 diabetes population (mean age 60.4 years). Of the population 44.9% was male, 96.9% had type 1 and 3.1% had type 2 diabetes (Abràmoff and Suttorp-Schulten, 2005). The images and their quality data were obtained in an anonymized manner in accordance with the Health Insurance Portability and Accountability Act (HIPAA) and the tenets of the declaration of Helsinki. The screening program accepts images from 20 different centers. The image size varied from 768×576 with 35° field of view (FOV) to 2048×1536 with 45° FOV. The diversity of the image data is shown by the example images in Fig. 2. All images were JPEG compressed. This lossy image compression format is a balance between image quality and capacity of data transmission and storage capabilities. The JPEG settings varied per screening site. Three types of fundus camera were used: the Topcon NW 100, the Topcon NW 200 and the Canon CR5-45NM. For this study all images were resampled, using cubic spline-based interpolation, such that their FOV were of approximately

equal size, namely 530 pixels in diameter. Images were acquired according to a fixed screening protocol. Two photos were obtained from each eye, one approximately optic disc (OD) centered and one approximately fovea centered.

Retinal images with a low quality should be reacquired, and therefore the image quality of each image is scored as part of the reading process. All readers (three in total) are ophthalmologists, with many years of experience in reading fundus photographs and diagnosing diabetic retinopathy. An image is scored as low quality when the ophthalmologist feels he or she is unable to provide a reliable judgement about the absence or presence of DR in the image. All other images are marked as normal quality. Independent training and test sets were created, each containing 1000 images (500 normal and 500 low quality). Each subject was represented in a set no more than two times, one image per eye. However, the images acquired from one subject are always either in the test or in the training set, never in both. Approximately 10% of all images in both the training and test set contained pathologies.

To compare the performance of the automatic method with that of a second observer, an ophthalmologist (MDA) assigned each image in the test set to one of four image quality categories. The four categories were; definitely low image quality, possibly low image quality, possibly normal image quality and definitely normal image quality.

3. Image structure clustering

Image Structure Clustering is a general way to learn the image structures that are present in a set of images. Provided with a previously unseen image the output of the ISC procedure gives information on the presence or absence of the image structures found in the training set. One of the possible applications of this technique is in the image quality verification of medical images, which is the goal of the present work.

The local image structure at each pixel can be described using the outputs of a set of filters. This set of filters (see for an example filterbank Fig. 3) generates response vectors of length l equal to the number of filters in the filterbank. The response vectors characterize particular image structures. For example, a filterbank consisting of first order Gaussian derivative filters at scale σ in the x and y direction ($l = 2$) will generate different response vectors at intensity edges with varying orientation. By using different filters and adding filters with varying scales σ, structures that exist at varying scales will generate specific response vectors. It is difficult to predict the responses of a filterbank to multi-scale, complex image structures. Therefore, it is important to start with an unbiased set of filters.

One could use histograms of the raw filter outputs over an image as features to characterize the presence or absence of certain image structures. Even with just a few bins per histogram the total number of features quickly becomes prohibitively large. Therefore, we propose to use an unsupervised clustering algorithm to cluster the response vectors into groups that characterize similar image structures in the image. This approach is similar to the one described in Malik et al. (2001) where filterbank responses at a certain scale are clustered to determine a set of texture primitives. In that work the authors are developing a framework for image segmentation. ISC attempts to find characteristic image structures, at multiple scales, in a set of similar medical images. This should lead to more specific clusters, partially corresponding to anatomical structures.

The characteristic structure clusters are found by applying the filterbank to a large amount of images and randomly sampling a number of response vectors from each image. Next, the resulting set of vectors is clustered using k-means clustering (Duda et al., 2001). Each cluster represent pixels that are on similar image structures, the image structure clusters. The final number of found clusters is equal to the parameter k of the clustering algorithm. Determining the optimal number of clusters for any given set is an unsolved problem and is therefore best determined empirically.

An unseen image is first filtered using the filterbank generating response vectors for all pixels in the image. The distance from the response vector to each cluster mean is measured and the pixel is assigned to the cluster with the nearest mean. In this way each pixel in an unseen image can now be assigned to one of the k clusters. The number of pixels in the image assigned to each cluster and their relative ratios then provide a compact description of the image structures in the image.

3.1. ISC for image quality verification in retinal images

For the specific application of ISC to the image quality verification of retinal images a number of choices had to be made. These encompassed the specific set of filters that should be used and their number as well as the total number of structure clusters. Important considerations are that when too many filters are applied, computational cost is increased as well as the correlation and redundancy between the filter outputs. A set of filters of limited size should be used, that can adequately describe the local image structure. The filters should be invariant to rotation and translation so as to generate similar responses for specific image structures regardless of their rotation or position in the image. This invariancy is important because one of the most important image structures in retinal images, the vasculature, can have many different orientations and can be located anywhere in the image.

Invariant filters based on first order gauge coordinates (ter Haar Romeny, 2003) satisfy these requirements. The gauge coordinate system is a coordinate system defined in each point of an image L by the first order derivatives. A local coordinate frame is defined $(\vec{v}, \vec{w})$ where $\vec{w}$ points in the direction of the gradient vector, $\left(\frac{\delta L}{\delta x}, \frac{\delta L}{\delta y}\right)$, and $\vec{v}$ is perpendicular to $\vec{w}$, thus pointing in the direction where the gradient is 0. As the gradient is invariant to rotation and translation, any derivatives expressed in gauge coordinates are invariants as well.

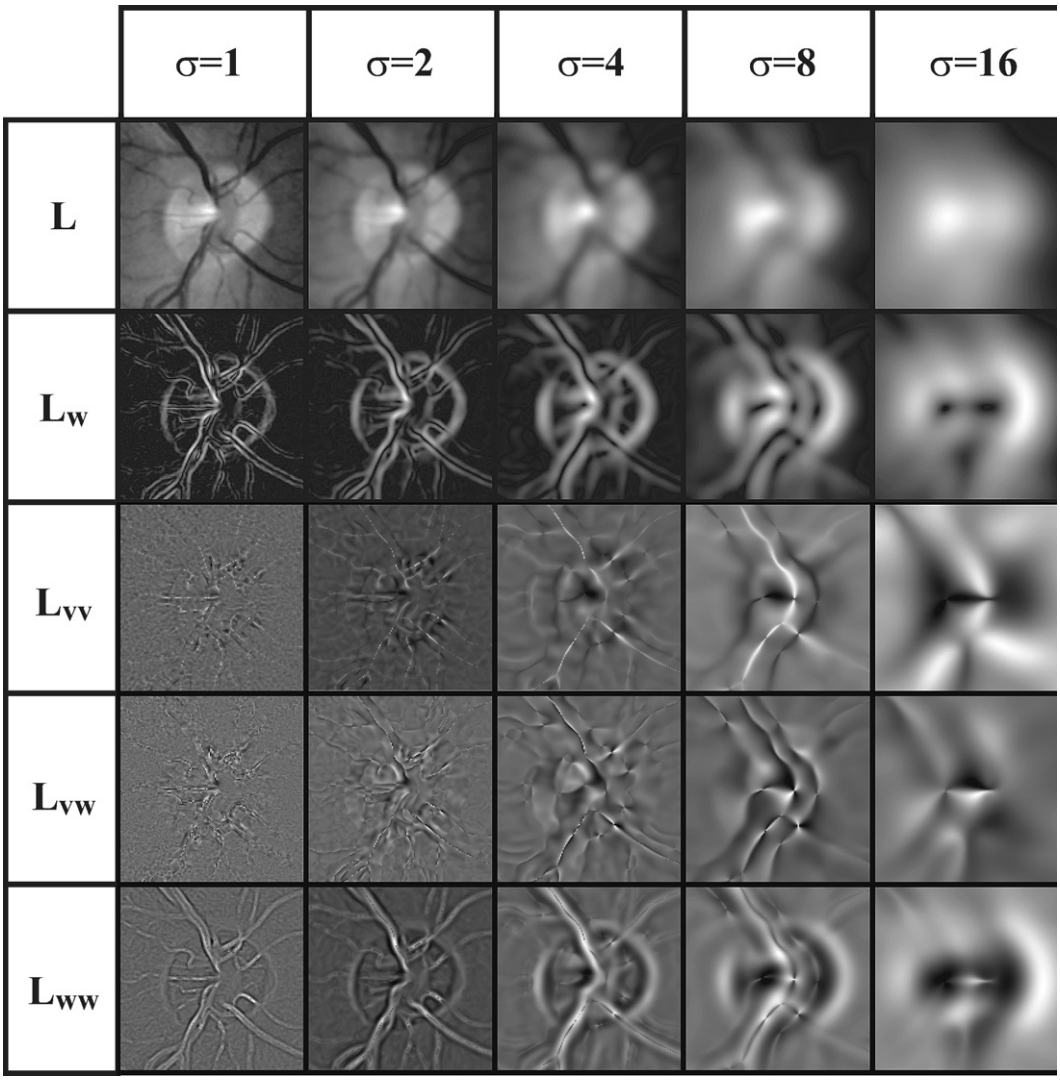

Fig. 3. Part of a retinal image, centered on the optic disc, convolved with all filters from the filterbank.

We used a second order irreducible set. Any other invariant up to and including second order can be expressed in terms of this set (ter Haar Romeny, 2003). Third and higher order invariants are not used as they are sensitive to noise and would substantially increase the total number of filters. The second order irreducible set consists of five filters; L, L_w, L_{vv}, L_{vw}, and L_{ww}. Here, the subscript indicates a derivative in a certain direction, e.g. L_{vv} is the second derivative in the direction of $\vec{v}$. The filters can be defined in image derivatives, i.e. L, L_x, L_y, L_{xx}, L_{xy}, L_{yy}, as shown in Table 1. Since Gaussian derivative filters are used to determine the image derivatives in x and y direction, structures that exist at different scales can be described by varying the filter scale parameter σ.

By visually examining the filter outputs at different scales we determined a set of five scales, i.e. $\sigma = 1, 2, 4, 8, 16$, that cover the scale range of normal image structures found in retinal images. Fig. 3 shows an image convolved with all filters in the filterbank. The filter kernels themselves are not shown as they are different depending on the local image structure (i.e. the gradient). Three color planes are available in retinal image data. Preliminary experiments showed us that including all the color information had a detrimental effect on the final output of the ISC procedure. Therefore, only the green plane image was used. The length of the response vectors of the filter bank thus was $5 \times 5 = 25$.

To characterize the image structures found in normal quality images filterbank response filters were generated

Table 1
The irreducible set of second order image structure invariants

Feature	Expression
L	L
L_w	$\sqrt{L_x^2 + L_y^2}$
L_{vv}	$\frac{-2L_xL_{xy}L_y+L_{xx}L_y^2+L_x^2L_{yy}}{L_x^2+L_y^2}$
L_{vw}	$\frac{-L_x^2L_{xy}+L_y^2L_{xy}+L_xL_y(L_{xx}-L_{yy})}{L_x^2+L_y^2}$
L_{ww}	$\frac{L_x^2L_{xx}+2L_xL_{xy}L_y+L_y^2L_{yy}}{L_x^2+L_y^2}$

The left column shows the set of irreducible invariants in gauge coordinates and the column on the right shows the invariants in x and y coordinates.

for every pixel in the set of 500 normal quality images from the training set. As there are a limited number of anatomical and image structures present in a retinal image, only a relatively small number of samples (i.e. response vectors) is needed to find the image structure clusters. From each image 150 response vectors were randomly sampled. In total 75,000 samples are collected that reside in a 25 dimensional space. All vectors were scaled to zero mean and unit variance and k-means clustering was applied. Through experimentation on the image data in the training set the number of clusters c giving the best classification performance was determined to be $c = 5$. In these experiments c was varied from 3 to 7 and the complete image quality verification system was then applied to subsets of the training set. For comparison, Fig. 4 shows the ISC output of the images in Fig. 1 with $c = 3$, $c = 5$, and $c = 7$.

In Fig. 4 every cluster has a different color. After visual examination of the results for $c = 5$, the following interpretation of each of the clusters can be given:

- Black: Background, dark to bright transitions.
- Blue: Background, bright to dark transitions.
- Green: Borders of high contrast structures on the retina and the fovea.

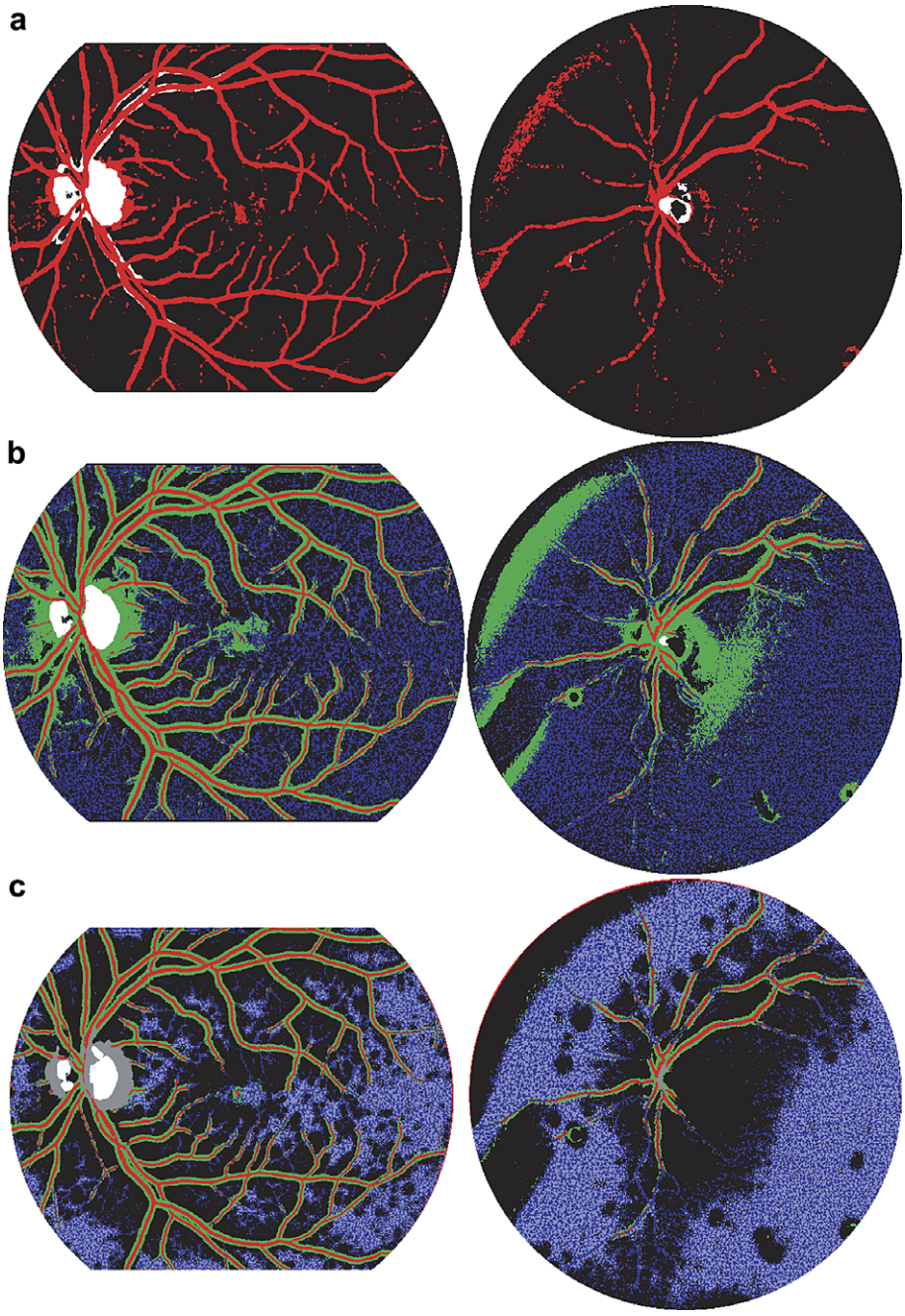

Fig. 4. Examples of the image structure clustering output of the images in Fig. 1. In each row the numbers of clusters c is different, $c = 3$ (a), $c = 5$ (b), and $c = 7$ (c). Each cluster is represented by a different color.

- Red: The vasculature.
- White: The optic disc.[3]

4. The complete image quality verification system

Our complete approach consists of three steps. First a set of features is extracted from the 1000 images in the training set. Then, feature selection is used to select the most salient feature sub-set. Finally, a classifier is trained using the selected features extracted from the training set. The trained classifier can then be directly applied to the test set to test its performance.

4.1. Features

The features used in the final system consist of a histogram of the ISC clustered pixels as well as the raw R, G, and B histograms. To produce the ISC histogram each cluster was assigned its own histogram bin. The RGB histogram features were added because they were often used in previous work. The histograms can help detect images with severe non-uniform illumination or low image contrast. The normalized histograms taken from the R, G, and B image planes formed the second part of the complete starting feature set.

To determine the number of bins for the RGB histogram features, histogram sizes of 5, 10, and 20 bins were tested on the training set. A classifier was trained using RGB histogram features exclusively and the best performance was found using histograms with 5 bins. The performance difference between 5 and 10 bins was small so the choice of the number of histogram bins for the RGB features does not appear to be a critical parameter.

The complete feature set is now as follows:

(1) The 5 bins of the normalized histogram of the image structure clusters.
(2) The 5 bins of the normalized histogram of the red image plane.
(3) The 5 bins of the normalized histogram of the green image plane.
(4) The 5 bins of the normalized histogram of the blue image plane.

In total 20 features are extracted for a single image.

4.2. Feature selection

Supervised systems often benefit from feature selection. Feature selection attempts to find a set of features, i.e. a sub-set of the complete feature set, that allow maximum separation of different classes of samples in the training data. In case of the proposed system, features that best separate normal from low quality images should be selected.

The sequential forward floating selection (SFFS) algorithm (Pudil et al., 1994) was used for feature selection. This is a wrapper-based algorithm, which means that it tests the classification performance of a specific classifier using different feature sets. Compared to other techniques, this algorithm has shown good performance on practical problems (Jain and Zongker, 1997). The algorithm employs a "plus 1, take away r" strategy. Features are added sequentially to an initially empty feature set but at every iteration features are also removed if that improves performance. In this way "nested" groups of good features can be found. For the feature selection experiments the training set was randomly divided in a feature selection training set, FS_{train}, and a feature selection test set, FS_{test}. Both sets contained 50% low and 50% normal quality images. The SFFS algorithm trained a classifier using the selected feature sub-set and FS_{train}, next the performance of the classifier and feature sub-set was tested on FS_{test}. The criterion used to measure the performance was the area under the ROC curve (Metz, 1986), A_z this measure lies between 0.5 for a method that is not better than random guessing and 1.0 for a perfect classification.

After feature selection had finished, two classifiers were trained. One with all features and one with the selected features. Whichever set of features showed the best classification performance on the training set was used in the final experiment.

4.3. Classifiers

It is difficult to predict beforehand which classifier will give the best performance for a particular classification task. Therefore, various classifiers classifiers were tested. These were a non-linear Support Vector Machine with radial basis kernel (SVM) (Chang and Lin, 2001), a Quadratic Discriminant Classifier (QDC) (Duda et al., 2001), a Linear Discriminant Classifier (LDC) (Duda et al., 2001) and a k-Nearest Neighbor Classifier (kNNC) (Arya et al., 1998).

For the kNNC the value of parameter k was determined using leave one out experiments on the training set using all features. The found value of k was then used in all experiments involving this classifier.

The SVM has a slow training phase in which, through cross-validation on the training set, the optimal values for its two parameters c and σ need to be determined. Parameter c is a penalty term for overlapping classes while parameter σ is the variance of the radial basis kernel used by the SVM. Both parameter values were found by using the "grid search" procedure proposed in Chang and Lin (2001). The fact that these parameters have to be determined make a wrapper based feature selection procedure, such as described previously, impractical. Good parameter settings need to be found for each tested feature combina-

[3] Colour images are only available in the online version of this paper.

tion. It is possible to use the features selected for another classifier in the SVM experiments. The features selected for the LDC should give a reasonable linear separation of the data. However, using these features the results on the training set showed a clear decrease in overall SVM classification performance for all system setups. Therefore, no feature selection was performed for the SVM.

5. Experiments and results

5.1. Experiments

The general algorithm for image quality verification has been outlined above. The two aspects of the algorithm that can be varied are the classifier and the feature set that is used. To show the effectiveness of the proposed ISC-based features and determine the optimal system configuration, different system setups, using different feature sets and different classifiers have been tested.

As far as the features are concerned, three basic system configurations were tested. A combination of ISC and RGB features, (*ISC + histogram*), only ISC features (*ISC*) and only RGB features (*histogram*). For each system feature selection was performed, as described earlier, and where this improved performance on the training set the selected features were used, otherwise the complete set of features was used. All features were always scaled to zero mean and unit standard deviation before feature selection and training.

For the final experiments on the test set each classifier was trained using the complete training set and the selected features. The trained classifier was then applied to the test set. After application of the classifier each image has been assigned a posterior probability representing the probability an image has normal quality. This process was done for all three systems and all classifiers resulting in a total of 12 performed experiments. In addition to the automatic systems, the performance of a second observer was also available (see Section 2).

5.2. Results

Classification performance in two class classification problems can be evaluated using ROC analysis. ROC curves plot the true positive fraction versus the false positive fraction. The ROCKIT software package (Metz et al., 1998) was used to produce the ROC curves. This software uses maximum likelihood estimation to fit a binomial ROC curve to the data. It also allows for statistical significance tests of the difference between ROC curves. In the case of the second observer three points on the ROC curve are known. All images were assigned to one of four classes where class 1 represents low quality and class 4 represents normal quality (see Table 3).

Table 2 shows the results of the different systems, using different classifiers. The results are given in area under the ROC curve and the accuracy (i.e. the number of correctly classified images divided by the total number of images). The posterior probability threshold at which the accuracy was calculated was the threshold at which the accuracy on the training set was maximal. The tables also show the parameters values for the SVM and kNN and whether feature selection was used in an experiment. ROC curves of the best performing system setups of *ISC + histogram*, *ISC* and *histogram* are given in Fig. 5 together with the results of the second observer. Area under the curve of the second observer is 0.9893(0.0035) with a 95% confidence interval (0.9803, 0.9944).

As *ISC + histogram* and the second observer are close to each other in terms of A_z, we performed a univariate z-score test of the difference between the areas under the two ROC curves. Here a p-value < 0.05 signifies there is

Table 2
Results of the three tested systems using different classifiers

Classifier	A_z	95% CI	Acc.
ISC + histogram			
SVM $c = 16384$, $\sigma = 9.77 \times 10^{-4}$	**0.9968(0.0013)**	**(0.9934, 0.9985)**	**0.974**
QDC*	0.9944(0.0014)	(0.9909, 0.9967)	0.963
LDC	0.9901(0.0021)	(0.9851, 0.9936)	0.951
kNN* $k = 15$	0.9932(0.0019)	(0.9885, 0.9961)	0.958
ISC			
SVM $c = 16$, $\sigma = 0.25$	0.9905(0.0021)	(0.9854, 0.9940)	0.953
QDC	**0.9948(0.0014)**	**(0.9914, 0.9969)**	**0.954**
LDC	0.9846(0.0029)	(0.9779, 0.9894)	0.936
kNN $k = 17$	0.9894(0.0024)	(0.9837, 0.9933)	0.950
Histogram			
SVM $c = 16384$, $\sigma = 3.91 \times 10^{-3}$	**0.9337(0.0074)**	**(0.9178, 0.9471)**	**0.860**
QDC*	0.9052(0.0092)	(0.8859, 0.9220)	0.825
LDC*	0.9049(0.0092)	(0.8855, 0.9217)	0.849
kNN* $k = 31$	0.9216(0.0082)	(0.9042, 0.9365)	0.844

A* behind the classifier name indicates a selected feature set was used. System performance is given in Area under the ROC, A_z, with standard deviation and 95% confidence interval (CI). The accuracy is in the rightmost column.

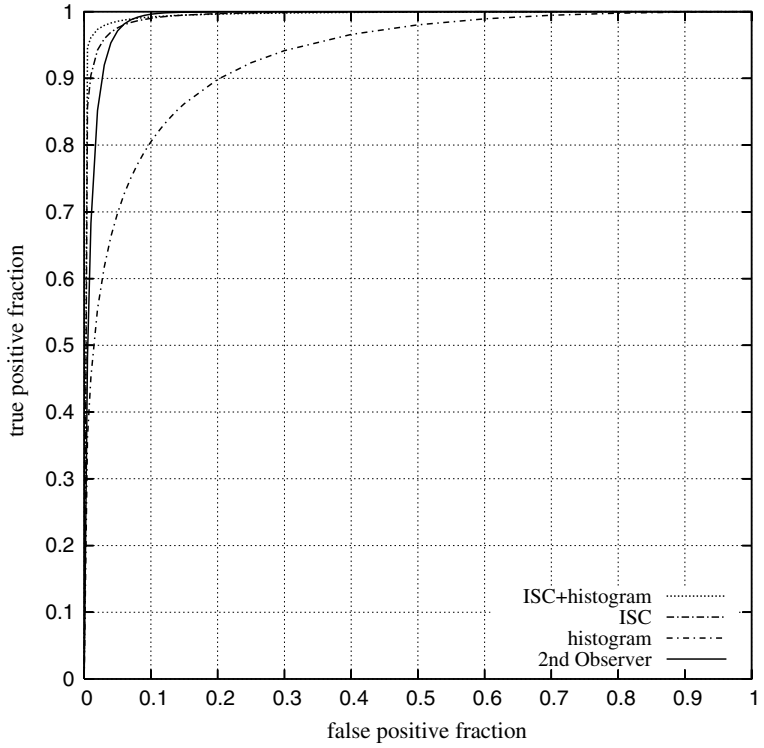

Fig. 5. The ROC curves of the three best performing versions of *ISC + histogram*, *ISC* and *histogram*. In addition to these, the ROC curve of the second observer is shown as well.

less than a 5% chance both datasets arose from binormal ROC curves with equal area under them. The difference between *ISC + histogram* and the second observer was found to be significant with a p-value of 0.019. A similar analysis was also performed for *ISC + histogram* and *ISC*, the difference between these two systems was not significant with a p-value of 0.069.

6. Discussion and conclusion

The image quality verification system presented in this paper shows excellent results. A number of system setups were tested to provide insight into the performance of the system using different feature sets and different classifiers. The best performing system, *ISC + histogram* with an SVM classifier and using no feature selection, achieved an area under the ROC of 0.9968 which is close to optimal. The experiments were conducted on a large, representative set of screening images. They showed that the ISC-based feature set was needed to attain good performance.

A statistical test on the A_z values of the best performing versions of *ISC + histogram* and *ISC* showed no statistically significant difference between the two. Nevertheless we prefer to use *ISC + histogram* in practice as it has a higher accuracy and the difference in A_z was almost significant with a p value of 0.069. In the following we will refer to this particular system as "the proposed system".

The main causes for the success of the ISC-based features lies in the fact that the clusters are based on data taken from a training set of similar, normal quality images. The results show that in low quality images the image structures and their relative ratios are indeed different. Examination of the results of the ISC procedure, see Fig. 4, show that the clusters that are formed capture the normal retinal anatomy such as the vasculature and the optic disc. The ability of the ISC output to describe the image structures in a compact fashion depends on the adherence to a protocol during image acquisition. A stricter adherence to the protocol should result in a more compact representation. If there is no protocol or the content of the images is very heterogeneous the effectiveness of ISC for compact image structure representation is reduced. In the work on textons (Malik et al., 2001) which deals with a heterogeneous set of images the number of used clusters is much larger, e.g. $c = 64$.

The most prominent structures appear first with less important structures added as the number of clusters is increased. The fovea does not seem to form a cluster, even with $c = 7$ clusters (see Fig. 4c). Probably this is due to its low contrast in many normal quality images in our set. The absence of certain structures in the image does not necessarily have to indicate the image has a low quality. Other causes could be heavy pathology or imaging of a non-standard part of the anatomy (i.e. image was not acquired according to the image acquisition protocol). Since automatic screening is our primary objective, and system sensitivity is extremely important in this situation, it is no problem if these images are also flagged as "low quality".

In a screening setting all low quality images should be examined by an ophthalmologist.

The ISC output could potentially be used for other purposes than image quality detection as well. One could use it as a feature to distinguish different types of medical images from each other automatically. Or it could be used to look for specific patterns in images, for example, in our application the different clusters could be used to find starting positions for segmentation algorithms of for instance the vasculature or the optic disc.

The fact that the best performing automatic system significantly outperforms the second observer may seem peculiar. As the original reference standard was also set by a human observer one would expect the second observer not to be significantly worse than an automatic system trained on the reference standard. However, although the difference

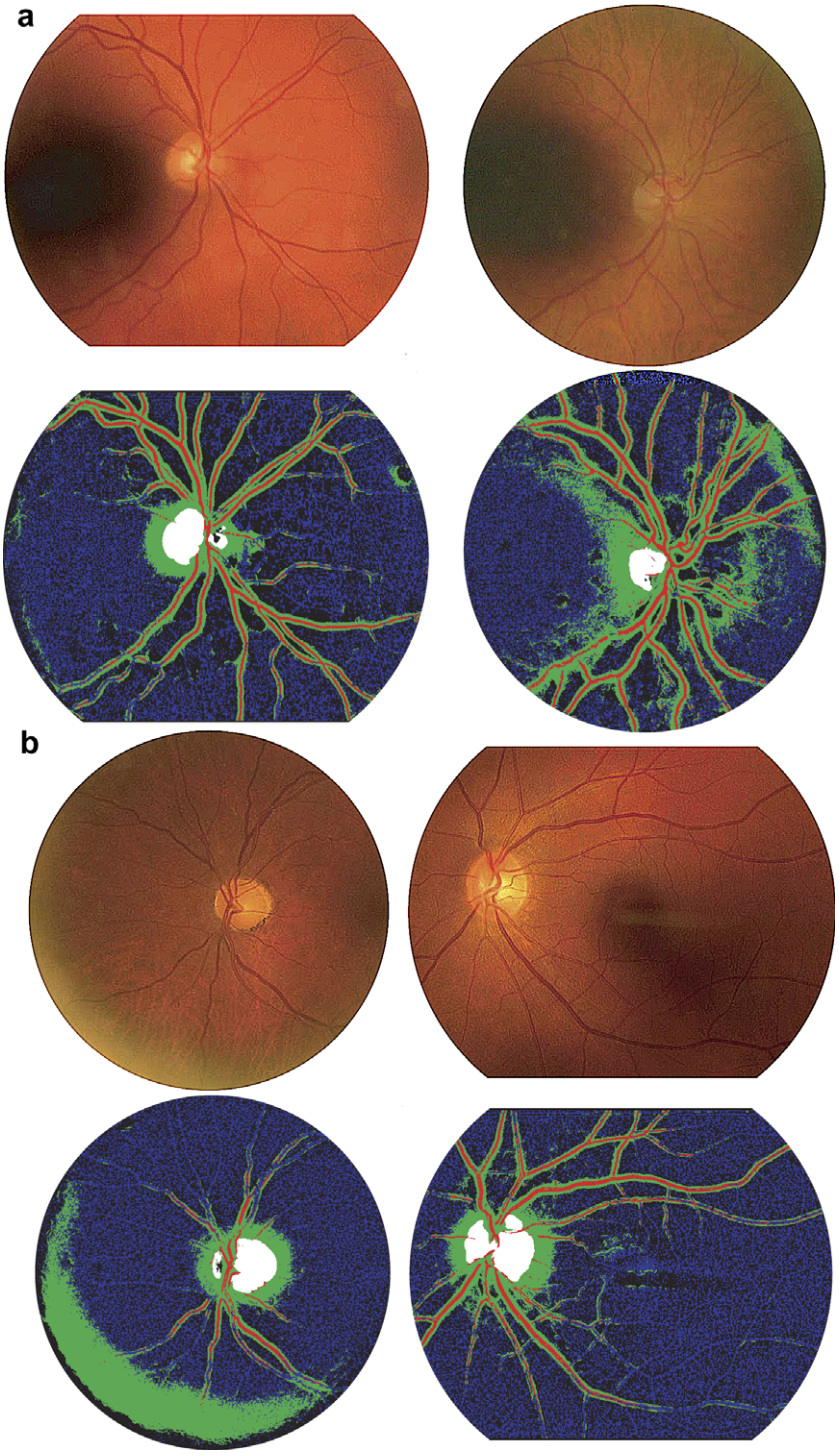

Fig. 6. Two examples of low quality retinal images and their ISC output that were incorrectly classified by the best performing system as normal quality (a), and the same for two normal quality images (b).

is significant it is small and might be explained by the way in which the second observer experiment was set up. To plot an ROC different operating points of an observer need to be known. Therefore, the observer was asked to assign each image in the test set to one of four classes. The original reference standard only gave one judgment regarding the image quality, either low or normal. The task assigned to the observer during the study was therefore different from the task of the original reader. When we look at the specific results given in Table 3 it is clear that the observer has no trouble finding the low quality images as 99.2% of the low quality images are assigned to the lowest two classes. For the normal quality images this is different, as only 67.5% of these images are assigned to the highest two classes. The remaining images are, except 2, all in class 2. This is an indication that had the observer been asked to split the set into normal and low quality images the results could have been different and the results would probably be different. It also indicates that there is a large variation in image quality amongst the normal quality images in the test set.

Although the best performing system comes close to the optimal A_z of 1, this point is not reached. The attained accuracy of 0.974 also shows there is still room for improvement. After examination of the false positive (FP) and false negative (FN) results produced by the system (see Fig. 6 for examples) we noticed that especially images that only exhibited low image quality in a localized area of the image were misclassified. This lead us to perform another experiment with a system that extracted all the same features as the best performing system but then separately from three different areas in the image. In theory this should allow the system to better detect images in which only a part of the image has low quality. However, the final performance of the system measured in A_z was slightly worse, although not significantly worse with a p value of 0.1389, than that of the best performing version of *ISC + histogram*.

A careful inspection of the cases shown in Fig. 6 allows us to speculate on what might have caused their misclassification. As we use global histograms as features in our system if the majority of the image has a good quality (i.e. the contrast is excellent) local problems can remain undetected (see Fig. 6a). A number of other cases in which this same problem occurred were correctly classified, there the RGB features probably allowed the system to detect the problem. In those cases a disproportionate number of pixels would have been in the lowest bin of the RGB histograms. Fig. 6b shows two normal quality images that were incorrectly classified. In these cases low contrast caused part of the vascular network to be missed and added to the background class. The left image also exhibits a large flash light artifact at the border of the FOV. This artifact didn't cause the ophthalmologists to mark this image as low quality however.

Table 3
Results of *Observer II*

Class	1	2	3	4
Normal quality images	2	161	122	215
Low quality images	448	48	1	3

Classes are numbered 1–4, with 1 signifying low quality and 4 signifying normal quality.

The total running time of the system on a new image is approximately 30 s. The software has not been optimized extensively and therefore further increases in speed can be expected.

To summarize, ISC allows for a compact representation of the structures found in an image. Features extracted from ISC output significantly improve the classification performance of an automatic system for the verification of retinal image quality. The system does not require any previous segmentation of the image in contrast with some previous work. Image quality is an important problem in large scale retinal screening for DR. As such the presented system may prove an essential part of a DR screening system.

References

Abràmoff, M., Suttorp-Schulten, M., 2005. Web-based screening for diabetic retinopathy in a primary care population: the Eye Check project. Telemedicine and e-Health 11 (6), 668–674.

Arya, S., Mount, D., Netanyahu, N., Silverman, R., Wu, A., 1998. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. Journal of the ACM 45 (6), 891–923.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification, 2nd ed. John Wiley and Sons, New York.

Fleming, A., Philip, S., Goatman, K., Olson, J., Sharp, P., 2006. Automated assessment of diabetic retinal image quality based on clarity and field definition. Investigative Opthalmology and Visual Sciences 47 (3), 1120–1125.

Jain, A., Zongker, D., 1997. Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2), 153–158.

Klonoff, D., Schwartz, D., 2000. An economic analysis of interventions for diabetes. Diabetes Care 23 (3), 390–404.

Lalonde, M., Gagnon, L., Boucher, M.-C., 2001. Automatic visual quality assessment in optical fundus images. In: Proceedings of Vision Interface 2001. Vision Interface.

Lee, S., Wang, Y., 1999. Automatic retinal image quality assessment and enhancement. In: Proceedings of SPIE Image Processing. SPIE Conference on Image Processing, pp. 1581–1590.

Lowell, J., Hunter, A., Habib, M., Steel, D., 2005. Automated Quantification of Fundus Image Quality. In: Proceedings of the 3rd European Medical and Biological Engineering Conference, pp. 1618 (1–5).

Malik, J., Belongie, S., Leung, T., Shi, J., 2001. Contour and texture analysis for image segmentation. International Journal of Computer Vision 43 (1), 7–27.

Metz, C., 1986. ROC Methodology in radiologic imaging. Investigative Radiology 21 (9), 720–733.

Metz, C., Herman, B., Roe, C., 1998. Statistical comparison of two ROC curve estimates obtained from partially-paired datasets. Medical Descision Making 18, 110.

Pudil, P., Novovicova, J., Kittler, J., 1994. Floating search methods in feature selection. Pattern Recognition Letters 15 (11), 1119–1125.

ter Haar Romeny, B., 2003. Front-End Vision and Multi-Scale Image Analysis, 1st ed. Springer, Dordrecht, The Netherlands.