

Vision-based, Real-time Retinal Image Quality Assessment

H. Davis¹, S.R. Russell², E.S. Barriga^{1,3}, M.D. Abramoff², and P. Soliz^{1,2}

¹*VisionQuest Biomedical*, ²*University of Iowa Department of Ophthalmology and Vision Sciences*, ³*University of New Mexico, Department of Electrical and Computer Engineering*
bert@visionquest-bio.com

Abstract— Real-time medical image quality is a critical requirement in a number of healthcare environments, including ophthalmology where studies suffer loss of data due to unusable (ungradeable) retinal images. Several published reports indicate that from 10% to 15% of images are rejected from studies due to image quality. With the transition of retinal photography to lesser trained individuals in clinics, image quality will suffer unless there is a means to assess the quality of an image in real-time and give the photographer recommendations for correcting technical errors in the acquisition of the photograph. The purpose of this research was to develop and test a methodology for evaluating a digital image from a fundus camera in real-time and giving the operator feedback as to the quality of the image. By providing real-time feedback to the photographer, corrective actions can be taken and loss of data or inconvenience to the patient eliminated. The methodology was tested against image quality as perceived by the ophthalmologist. We successfully applied our methodology on over 2,000 images from four different cameras acquired through dilated and undilated imaging conditions. We showed that the technique was equally effective on uncompressed and compressed (JPEG) images. We achieved a 100 percent sensitivity and 96 percent specificity in identifying “rejected” images.

Index Terms — Image Quality, retina, ophthalmology

I. INTRODUCTION

The objective of this research was to implement a system that will enable real-time quality assessment of any color fundus image taken from any digital camera. This research discovered the individual key image parameters that aid in determining the visually classified quality characteristics of a retinal image. Our goals were to: 1) determine the quality of an image within milliseconds (real time); 2) perform image quality assessment with high sensitivity and very high specificity (99% and 95%, respectively); and 3) identify sources of image quality to assist the photographer in making appropriate corrections.

II. BACKGROUND

Obtaining the highest possible image quality is critically important when photographing a patient’s retina in a clinic or collecting images of a subject for a study or clinical trial. Often the photographer taking the image will not appreciate the requisite criterion for image quality required by

the end user, whether an ophthalmologist or a highly trained grader or reader of the retinal images. What may appear acceptable to the photographer may be deemed unacceptable or entirely unusable by the grader or clinician. In telemedicine, transmitting an unacceptable quality image may mean, at worse a missed diagnosis, or at best the need to retake the image or images at an inconvenience to a patient who will have to return to the clinic for re-imaging. In longitudinal studies where every examination and image is critically important, losing an image for a given examination period may result in loss of that individual from the study or a reduction in the statistical power of the results.

Zimmer-Galler and Zeimer [1] found that in a multi-site study of 2771 patients where 304 (11%) of the images were found unreadable, approximately 25% were due to poor patient fixation, 25% due to poor focus, 25% due to small pupil size or media opacity. The remaining cause(s) for unreadable images was undeterminable. We describe in this paper an image quality system that detects poor quality images which are a result of these factors.

High-quality images are a prerequisite for automatic screening systems, such as those for diabetic retinopathy. With digital systems, it is possible to implement a “Real-time Image Quality” grading system where the personal computer or laptop storing and/or transmitting the images can automatically and immediately (one or two seconds after taking the image) indicate to the photographer an image quality score, such as “good – acceptable – unacceptable.”

It is evident that human visual perception should be at the core of quantitative and automatic image quality evaluation system. Image quality assessment is a challenging task that has traditionally been advanced by computational models based on a systematic and computer-oriented approach to performing a series of subtasks leading to the completion of the image quality assessment. Investigators have addressed the problem of automatic or quantitative image quality assessment of digital images in a variety of ways. From the broader need for image quality assessment of color photographs to the specific need for optimum image quality of medical images from various modalities (radiographs, computer tomography, ultrasound, magnetic resonance imaging, etc.), there has been an interest in developing computer-based image quality assessments tools.

Image quality issues can be divided into four general areas: Physics, Observer, Task Requirements, Patient Effects. Ultimately, an automatic image quality system must consider

each of these issues and provide the photographer or user with diagnostics as to why an image may be less than optimal. Each of these effects is considered explicitly in our image quality analysis and coupled directly to the human grader-based gold standard for perceived image quality. In other words, our image quality model will mimic the human graders since the model is designed to “fit” the human graders’ image quality standards.

III. RELATED RESEARCH

Automatic image quality assessment has been the topic of intense study by a number of researchers in other fields of medicine as well as the general topic of image quality. Wang and Bovik, who have been leaders in the study of image quality, have published several seminal papers and book chapters on the topic [2, 3, 4, 5, 6]. One observation from the study of Bovik’s body of research on image quality is that image quality assessment should involve the human observer, as does our methodology. Most image quality models are based on a comparison to a reference image and are reflected in much of Bovik’s work. Using a reference image, quality comparisons can then be made using various quantitative measures. A reference image (the optimal image of each retina) is not available in our application. Nevertheless, their studies are highly relevant and have influenced the approach and direction of this research.

The automatic detection of retina image quality has received attention in ophthalmology. Lee and Wang [7] introduced the idea of using edge intensity histograms to characterize the sharpness of the image. By itself this approach did not give totally satisfactory results. As Lalonde [8] states, “The link between intensity histogram similarity and image quality is not that strong.” Others, like Fleming et al. [9], Usher [10], and Damera-Venkata et al. [11] have applied techniques that involved segmentation of retinal vessels and other anatomical structures, then applying additional criteria, such as detecting small vessels around the fovea, to determine image quality.

The presence of small vessels in this area is used as an indicator of image quality. For example, Fleming, et al. [9] have developed a computational model for evaluating the acceptability of retinal images based on several criteria related to image quality. Several steps are required, including segmenting retinal vessels. To evaluate the clarity of an image, retinal vessels in the vicinity of the fovea are counted and measured. The clear presence of these vessels will indicate a high-quality image.

Niemeijer, Abramoff, et al. have presented a related method called Image Structure Clustering (ISC) to provide a compact representation of the structures found in an image [12]. Their 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 filter bank.

Other factors that affect image quality are individually addressed and quantitative criteria are set for each. Though the results are good, as was demonstrated by Fleming, this technique is computationally burdensome and must integrate explicitly all possible factors, each treated independently with a different algorithm. 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

the segmentation in case of low image quality is not trivial and limits the robustness of the approach suggested by Fleming and other investigators.

IV. FEATURE-BASED IMAGE QUALITY

This paper presents a method that is based on computationally simple features. For each color channel in the RGB and CIELab image, a total of $N = 17$ features were calculated. The features were coded for the CIELab space.

Table 1 – Features used for determining image quality.

Mean Intensity	Variance
Skewness	Kurtosis
Entropy	8 Haralick [13] features
Spatial frequency	1 st Quartile
Median	3 rd Quartile

These features were produced for each of the three color spaces. The feature set was produced for each of seven regions of the retinal image in order to obtain the effects of spatial variations on image quality. The two models resulted in features for each of the seven regions. All features were calculated for the active areas of the image (Figure 1).

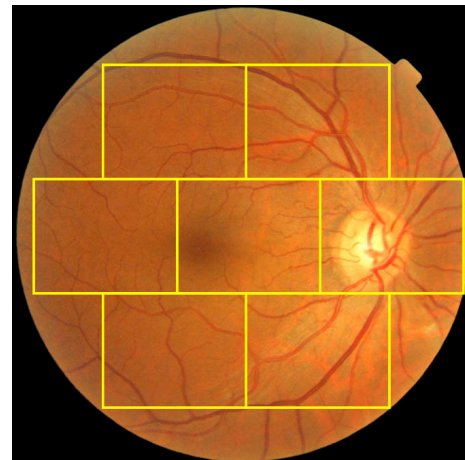


Figure 1 – A grid is placed showing the regions where the features are calculated.

Color Model and Visual Perception

Many studies have used the three color channels, red (R), green (G), and blue (B) directly. A color model that represents the color perception qualities of the human eye are best represented by CIELAB or the derivative CIE $L^*a^*b^*$ model. (The name comes from the Commission Internationale d’Eclairage, an organization that developed the model.) In this representation, colors values are linearized with respect to perceptual color differences. This means that a change in a measured color value will produce the same relative change in the visual properties. This approximation to the human vision system will improve the discovery of the relationship between the objectively derived image features and the graders’ response to an image’s quality. Conversion equations are found in most image processing texts, such as Gonzales and Woods [14].

Luminance Features for Image Quality Correspondence Vector

A standard digital color image has three color planes (red, green, and blue) that, when combined digitally, represent the various hues and intensities or colors observed in an image. In retinal photographs, the red channel is always the brightest, while the blue channel has the lowest intensity values. This is due to several reasons including 1) the sensitivity of the digital cameras, which are most sensitive to red light, and 2) the stronger absorption of blue light by the anterior segment of the eye and stronger reflectance of red light by the posterior segment of the eye. This suggests that the human observers, whether ophthalmologists or trained ophthalmic technicians, will develop a preference for retinal photographs with certain balance between the three color channels [15]. On this basis, the first set of features will be ratios of the average intensity of the three color channels.

Feature A: μ_L , μ_a , μ_b are defined as the mean intensities for each channel. As demonstrated above, low or high exposure will be a factor in reducing image quality.

Feature B: Skewness for each color channel. Skewness is a measure of symmetry, or more precisely, the lack thereof.

Feature C: Kurtosis for each color channel. Kurtosis is a measure of whether the histogram of pixel intensities is peaked or flat relative to a normal distribution.

These features will provide a complete characterization of the luminance in the image.

Contrast Features for Image Quality Correspondence Vector

Feature D: The variance of the intensity within each of the regions in Figure 1 and for each color channel will be another set of individual color channel features. These features will characterize quality aspects of the image that indirectly represent the contrast in the image. Low variance will reflect low contrast, regardless of overall image brightness.

Feature E: Co-occurrence contrast from Haralick features. This metric explores the relationship of a pixel to other pixels in its neighborhood. A distribution of the difference in gray-level intensities between these pixels is used to calculate a contrast value.

Feature F: Entropy from Haralick texture is a measure of the representation of quantized gray levels. An even distribution (high contrast) will have the largest possible entropy value.

Feature I: Spatial frequency is affected by both contrast and noise. Sharp edges, such as those produced by retinal vessels, will increase the spatial frequency; however, so will shot (speckle) noise. For this reason alone, spatial frequency by itself will not likely produce an image quality index.

V. ANALYTICAL METHODS

Partial least squares (PLS) [16, 17, 18] was used to develop a predictor of image quality. The images were given a score by the grader and quantized to a numerical value for use by PLS. This score was used as the dependent variable, and the above features were used as the independent predictor variables. PLS

was then used to produce a linear regression equation which gives a continuous score for quality.

PLS is a very powerful method of eliciting the relationship between one or more dependent variables and a set of independent (predictor) variables. Numerous other methods exist to predict quality given a set of independent variables. There are certain circumstances, however, where some will fail to work. One is multicollinearity, a condition that occurs when some of the independent variables are highly correlated with each other. In the present circumstance, the same feature when applied to different color channels tends to be highly correlated even if their magnitudes are quite different. Hence, the quality assurance problem using a common feature set is highly multicollinear.

Experiment #1: MESSIDOR retinal image data

To develop and test the image quality algorithms, we first selected $N = 600$ from a diabetic retinopathy research database, MESSIDOR (Available on line at <http://messidor.crihan.fr/>). The digital color retinal images were captured using 8 bits per color plane and a resolution of 2240 x 1488 pixels. Images were acquired using a Topcon TRC NW6 with a 45 degree field of view. The images were taken with pupil dilation and were stored in an uncompressed TIF format.

We randomly selected right and left eye retinal images ($N=200$) from the database, which were of varying image quality (0 to 5, where 0 is low quality and 5 is high quality). This database did not contain retinal images that would be judged to be ungradeable. Because these images are generally good quality, to perform initial testing on the algorithms, we artificially degraded the images by applying Gaussian filters to blur 200 images. We also reduced or increased overall image intensity of 200 other images by shifting each of the color channel's histogram to the right or left to darken or lighten, respectively. This simulated rejected image quality was based on analysis of the characteristics of actual images which had been rejected by ophthalmologists in the Iowa database. Figure 2a shows a sample "acceptable" image, with a blurred image in Figure 2b and a over exposed image in Figure 2c. These first sets of results represent our first experiment to test the algorithms' performance in detecting retinal images with unacceptably low quality, i.e. *rejects*.

Training and testing was performed using the jackknife approach, where training is performed on all but one of the images and testing on the remaining image ("leave one out"). For the images in the remaining experiments, tests were performed on independent data sets.

The results of the first experiment that applied only nine CIELab features and the seven regions of the image to the Messidor database was that all rejected images were correctly identified by our method. Accuracy was 100%.

The PLS technique produces a set of factors that contribute most to the classification, i.e. feature weights. This allows one to assess the factors that are likely associated with the visual perception of image quality. Table 2 shows the percent of the total weight contributed by each feature the first ten features. Note that the 10th feature, "Magenta(a)*Q1", contributes only to 4% of the weight. 90% of the weight is accounted for by the first 15 features. This finding gives us the option to eliminate the less important features.

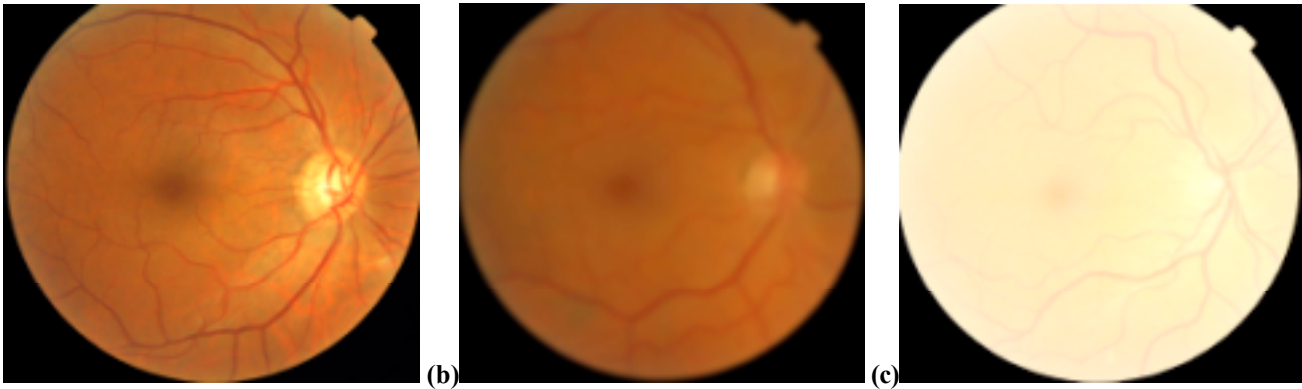


Figure 2. a) Good quality image as labeled in the Messidor database and as determined by Grader, b) Unfocused or blurry degraded image, c) overexposed image. Images b and c were graded as rejects by a trained ophthalmic technician. Underexposed images (not shown) were also part of the ‘rejected’ image set.

Table 2 – Experiment #1 – Fifteen most important factors.

#	Scale	Feature	%	Cum %
1	‘b’	SpFreq	15.9	15.9
2	‘a’	SpFreq	14.9	30.8
3	‘b’	Entropy	10.9	41.7
4	L	SpFreq	9.9	51.6
5	‘b’	Skewness	7.6	59.2
6	‘a’	StdDev	4.6	63.8
7	‘a’	Q3	4.3	68.1
8	‘b’	Q3	4.2	72.3
9	‘b’	Q1	4.0	76.3
10	‘a’	Q1	3.4	79.7
11	‘a’	Mean	2.5	82.2
12	L	Kurtosis	2.4	84.6
13	L	Entropy	2.1	86.7
14	‘a’	Skewness	1.9	88.6
15	‘b’	Mean	1.8	90.4

The results of our analysis show that the rejected images are differentiated from the acceptable images by the features shown in Table 2. In the CIELab color model, the “L” represents luminescence; “a” is the magenta contrast; and “b” is the yellow contrast. These are the variables to which the red, green, and blue of the RGB model have been mapped. Note that the “b” or yellow contrast appears five times in the top ten weighted features and contributes to 48% of the overall weight. “L” or luminescence contributes 17% and “a” or magenta contrast is 35%. The two most important features for determining image quality in this model are “b” spatial frequency and “a” spatial frequency. “L” spatial frequency follows as the fourth greatest weighted-feature. Spatial frequency is encoding the variations or undulations of the structures in the two contrast channels. “b” entropy, which is a measure of the distribution of gray-levels or loosely interpreted as the contrast of the channel contributes to about 12% of the total weight. This analysis was performed to eliminate features that contribute little to the classification.

However, because even with the processing of the several hundred features and classifying the image takes less than a second on a standard PC laptop computer, it is not necessary at this point to eliminate the low ranking features. In the future (Phase II) larger images and more regions may necessitate reducing the feature set.

An additional experiment (Table 3) was performed to determine how well our methodology could determine not

only that an image was below an acceptable quality, but suggest what the nature of the reduced image quality is; for example, over/under exposure or focus error. This determination would give the photographer a recommended action to try to correct the image acquisition. In Table 3, we show that we are able to classify at >99% accuracy the three categories: blur (out of focus or anterior segment opacities), over exposure, and under exposure. We envision future cameras or image handling systems with a capability to give photographers immediate feedback as to the technical basis for a rejected image, such as exposure or de-focus. The real-time retinal image quality assessment system would assist the photographer help provide training through drop-down windows and help menus.

Table 3. PLS prediction of good quality images versus blurred and versus over or underexposed images of unacceptable quality. Errors: one image was classified as acceptable which the grader called a bad exposure (out of 600 cases). One image was classified as blurred while the grader called it a bad exposure. Two images were classified as bad exposure, which the grader called blurred. Of the 600 images, there was only one false negative (classified as acceptable by PLS, but called unacceptable by the grader). Kappa = 0.998.

	Acceptable	Blurred	Bad Exposure
Acceptable	200	0	0
Blurred	0	198	2
Bad Exposure	1	1	198

Experiment #2

In this experiment, we selected 398 images from the University of Iowa database. Approximately half (52%) of the images ($N = 200$) were of acceptable quality, while ($N = 198$) were classified by an ophthalmologist as ungradeable. Figure 3 presents eight examples of the images found in this dataset. These images were not simulated or staged. These are images that were transmitted to the University of Iowa for purpose of diabetic retinopathy screening. The images **c** through **h** were deemed ungradeable by the ophthalmologists.

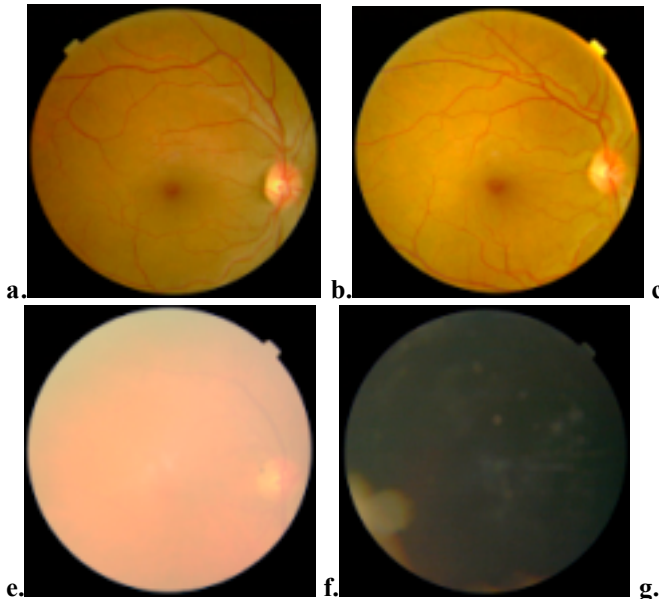


Figure 3. a) Good quality image as labeled in the database; b) Acceptable quality image though a minor lighting artifact; c) Image out of focus; d) Image with a lighting artifact; e) Underexposed; f) Underexposed; g) Underexposed; h) Underexposed with a lighting artifact.

The Data

These images were provided by Dr. M. Abramoff (MD, PhD) at the University of Iowa. Personal identifiers had been removed from this research database. The data falls under NIH exemption #4. The images are taken from twenty different sources in The Netherlands and Iowa. The diversity of the data format was a factor in its selection for this study in order to demonstrate the broad applicability of our methodology. Image sizes included: 2048 x 1536 pixels; 1792 x 1184; and 768 x 576. Field of view was either 30 or 45 degrees. All images were JPEG compressed from 10:1 to 20:1. The lossy image compression is a factor that must be addressed in a truly robust real-time retinal image quality assessment system. Three different cameras were used: Topcon NW 100, Topcon NW200, and the Canon CR5-45NM. The imaging protocol called for two fields (the first centered approximately on the optic disc and the second centered approximately on the fovea) for each eye. From the database of approximately 5,000 cases, we selected two sets of images. For this experiment (#2), we focused on developing a model to be tested in future research with an entirely different set of retinal images.

The Results

This experiment used 200 of the 398 cases. The 200 cases were selected randomly. All rejected images were identified with only a 4% false positive rate (7 acceptable quality images were misidentified as ungradeable). This experiment was repeated five times, each time with a different combination of 200 acceptable and rejected images. An average area under the ROC curve for the five tests was 0.993, indicating near perfect accuracy in identifying rejected images.

Figure 4 shows the relative weights of the top 16 features. The magenta contrast channel, “a”, contributes to 51% of the total

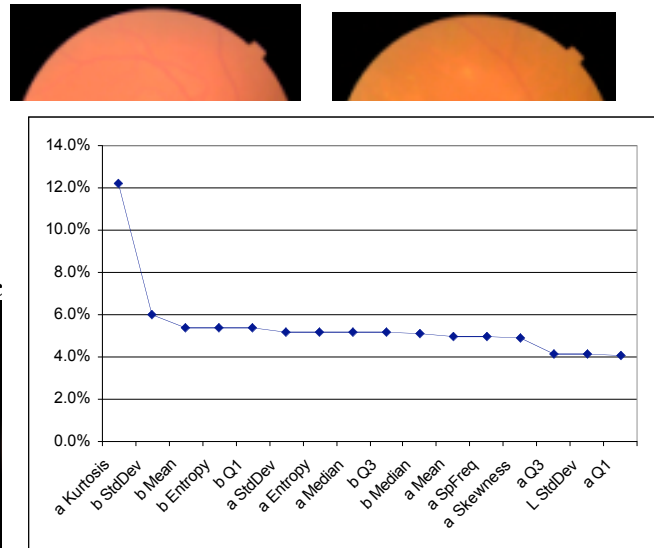


Figure 4. Weights of the top 16 features toward classifying reject versus acceptable images. Data points are color coded by red, green, or blue for the three color channels. Circled features exhibit the greatest weight.

weight. The “a” kurtosis statistic had the greatest magnitude at 12 %. The yellow contrast channel, “b”, contributes to 36% of the weight in the image quality classification model. 17% of the weight comes from the luminescence channel, “L”. Although the specific features are different from Experiment #1 (Messidor database), in both cases the magenta contrast channels dominates the weights. Similarly, in both experiments, the yellow contrast channel has the second highest weight, while the luminescence channel provides the least weight toward the image quality classification model.

VI. CONCLUSIONS

This research shows that using common image processing features, one can find those few that will robustly identify those images that are below a visually determined threshold for acceptable image quality. Future research will evaluate these techniques for finding similar thresholds for recognizable differences based on the amount of compression for an image.

ACKNOWLEDGMENTS

This research was funded by the National Eye Institute, EY018971. We also wish to thank the University of Iowa for

providing part of the data and the visual evaluation of the images for quality.

REFERENCES

- 1 Zimmer-Galler, I. and R Zeimer. "Results of Implementation of the DigiScope for Diabetic Retinopathy Assessment in The Primary Care Environment," Telemedicine and e-Health, Vol. 12, No. 2: 89 -98. Apr 2006.
- 2 Wang, Z, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004. Handbook of Image and 32 Video Processing
3. Wang, Z and A. C. Bovik, "A universal image quality index," IEEE Signal Processing Letters, vol. 9, no. 3, pp. 81-84, March 2002.
- 4 Wang, Z, A. C. Bovik and L. Lu, "Why is image quality assessment so difficult?" Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Proc., vol. 4, pp. 3313-3316, May 2002.
- 5 Wang, Z and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," Human Vision and Electronic Imaging IX, Proc. SPIE, vol. 5292, (San Jose, Jan 2004).
- 6 Wang, Z, E. P. Simoncelli and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," 37th IEEE Asilomar Conference on Signals, Systems and Computers. (Pacific Grove, Nov 2003).
- 7 Lee SC and Y Wang, "Automatic retinal image quality assessment and enhancement," Proceedings of SPIE Medical Imaging Processing, 3661:1581-1590. SPIE (Washington, DC 1999).
- 8 Lalonde, M, L Gagon, and MC Boucher, "Automatic visual quality assessment in optical fundus images," Proceedings of Vision Interface 2001, Ottawa, 259-264, June 2001. <http://www.cipprs.org/vi2001/schedulefinal.html>.
- 9 Fleming, AD, S Philip, KA Goatman, JA Olson, and PF Sharp, "Automated Assessment of Diabetic Retinal Image Quality Based on Clarity and Field Definition," Investigative Ophthalmology and Visual Science.47:1120-1125. 2006.
- 10 Usher DB, Himaga M, Dumskyj MJ, et al. "Automated assessment of digital fundus image quality using detected vessel area. Proceedings of Medical Image Understanding and Analysis." 2003;81-84. British Machine Vision Association (BMVA) Sheffield, UK:
- 11 Damera-Venkata, N, TD Kite, WS Geisler, BL Evans, A Novik, "Image quality based on degradation model," IEEE Trans. Image Processing, 9(4):636-650, 2000.
- 12 Niemeijer, M, B van Ginneken, and MD Abramoff, "Image Structure Clustering for Image Quality Verification of Color Retina Images in Diabetic Retinopathy Screening," Medical Image Analysis 10(6): 888-898, 2006
- 13 Haralick, R.M., Shanmugan, K., and Dinstein, I. 1973. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 3, No. 6, pp. 610-621.
- 14 RC Gonzalez, and RE Woods, Digital Image Processing. Addison Wesley, US, 1992.
- 15 Hubbard, LD, Danis RP, Neider MW, Thayer DW, Wabers HD, White JK, Pugliese AJ, and Pugliese MF, "Brightness, contrast, and color balance of digital vs. film retinal images in the Age-Related Eye Disease Study 2," Investigative Ophthalmology and Visual Sciences, Aug 2008.
- 16 Wold, H. "Personal memories of the early PLS development." Chemometrics and Intelligent Laboratory Systems, 58:83-84, 2001.
- 17 Wold, H.. "Estimation of principal components and related models by iterative least squares." in: P. R. Krishnaiah (Ed.), Multivariate Analysis. Academic Press, New York, 1966
- 18 Wold, H. "Path models with latent variables: the NIPALS approach" in: H. M. Blalock (Ed.), Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building. Academic Press, New york, 1975.