33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

# An Automated Retinal Image Quality Grading Algorithm

Andrew Hunter, James A. Lowell, Maged Habib, Bob Ryder, Ansu Basu, David Steel

*Abstract*—**This paper introduces an algorithm for the automated assessment of retinal fundus image quality grade. Retinal image quality grading assesses whether the quality of the image is sufficient to allow diagnostic procedures to be applied. Automated quality analysis is an important pre-processing step in algorithmic diagnosis, as it is necessary to ensure that images are sufficiently clear to allow pathologies to be visible. The algorithm is based on standard recommendations for quality analysis by human screeners, examining the clarity of retinal vessels within the macula region. An evaluation against a reference standard data-set is given; it is shown that the algorithm's performance correlates closely with that of clinicians manually grading image quality.**

## I. Introduction

THE clarity of fundus images used for diagnosis is an important issue in automated detection of retinal conditions such as diabetic retinopathy. Inadequate quality can affect subsequent diagnostic stages as quite subtle visual differences may distinguish diagnostic features such as retinal lesions from confounds (e.g. pigmentation variations). In more extreme cases, important features may blend completely into the retinal background. Ungradeable images should therefore be flagged for either repeat screening or ophthalmic review.

Digital fundal image quality can be affected by a number of factors including patient head or eye movement, poorly dilated and/or small pupils, blinking, and media opacity (e.g. cataract). Head or eye movement can result in out of focus, incorrectly illuminated or misaligned images. Retinal screening protocols require carefully aligned images to include defined areas of the retina. In a 45° macula centred image, protocol requires the optic nerve head to be positioned in the midline, one disc diameter from the edge of the image field. Any movement just prior to acquisition can cause misalignment and to vital regions being excluded from the photograph. Poorly dilated pupils may affect image illumination creating dark low contrast images and can prevent lesion identification. If fundal cameras capture retinal images through cataract, images appear blurred and are often ungradeable. Images can also be obscured by eyelashes or the eyelid if blinking occurs during acquisition.

Currently, assessment of image quality in the UK is based on subjective interpretation of three definitions of image clarity (adopted by The National Screening Committee (NSC) [1]). These guidelines suggest using two 45 degree field images per eye; image quality is based on a macula centred images. The three standard defined levels are:

Figure 1: Sample image of NSC "Achievable" Standard

Figure 2: Sample Image of NSC "Minimum" Standard

Andrew Hunter is at the University of Lincoln, UK. (ahunter@lincoln.ac.uk).

James Lowell is working in the image processing software industry.

Maged Habib and David Steel are at Sunderland Eye Infirmary, UK, (dhwsteel@hotmail.com).

Bob Ryder and Ansu Basu are at City Hospital, Birmingham, UK.

1. *Achievable standard*: Optic disc less than or equal to one disc diameter (1DD) from the defined position. Small vessels clearly visible within 1DD of the fovea and optic disc and visible across more than 90% of remaining image(s); see figure 1.
2. *Minimum standard*: Optic disc less than or equal to 2DD from the defined position. Small vessels clearly visible within 1DD of fovea and optic disc and visible across more than 66% of remaining image(s); see figure 2.
3. *Inadequate (ungradeable)*: Optic disc less than or equal to 2DD from the defined position. Small vessels not clearly visible within 1DD of fovea and optic disc and visible across more than 33% of remaining image(s); see figure 3.

Figure 3: NSC "inadequate" quality image

## II. QUALITY ASSESSMENT ALGORITHMS

Image quality measures are well-known in the domain of image restoration; however, diagnostic suitability is a relativity new research area with only limited publications. Lee *et al.* [2] studied 360 retinal images from the Oklahoma Native Americans and concluded that image quality could be defined by three parameters: brightness, contrast and signal-to-noise ratio (SNR). From the sample set, twenty images with excellent quality were selected. Quality parameters were obtained from these images together with an average intensity histogram – referred to as the desired values and the template intensity histogram respectively. Lee observed that the brightness, contrast and signal-to-noise values of an image were close to their respective desired values when the image's intensity histogram was close to the template intensity histogram and that these values could be derived from the histogram. An image quality measure was therefore proposed using the convolution of the template histogram with the image histogram and computing a quality index.

In an evaluation of Lee's work, Lalonde *et al.* [3] examined the interdependency between image quality and histogram similarity in 40 retinal images of varying quality. Histograms from several poor quality images were found to closely resemble the template histogram. In addition, histograms from several good quality images were notably different from the template histogram, signifying a weak connection between image quality and histogram similarity. Lalonde experimented with distribution of edge magnitudes and the local distribution (as opposed to the global histogram of Lee) of pixel intensity as quality indicators. In a similar approach to Lee, a typical edge magnitude histogram was formed using the edge maps from a set of good quality images. The difference between the typical and current image edge magnitude histogram formed a quality indicator. A second quality indicator was derived by comparing local intensity distributions. This approach differs from Lee *et al*, by defining a set of local histogram templates instead of one global histogram template. Lalonde concluded that both quality indicators could help discriminate between good and bad images, although a larger image set was required to evaluate the performance of the approach.

Niemeyer et al. [4] apply a set of filter banks to perform a coarse segmentation of the entire retina (into background, ONH, vasculature and high contrast edges), representing this as a histogram; they also construct 5-bin histograms of the R,G and B channels. They then trained an SVM classifier to distinguish between low and high quality images, reporting better results than Lee and Lalonde via this more sophisticated approach to histogram analysis which, importantly, takes into account the proportion of the image occupied by the vasculature and the image contrast – both key indicators of gradeability.

Usher *et al* [5,6] presented an algorithm that uses a quality metric based on the area of automatically detected blood vessel. Vascular segmentation is performed using a combination of orientated matched filtering and region growing, Usher reports a sensitivity of 81% and a specificity of 91%. As blood vessels should be present in all retinal images regardless of ethnic origin or retinopathy, Usher measured vessel frequency from gradeable and ungradeable images to determine an image quality metric. Within each image an image quality metric score $V$ was set from the total count of pixels classified as vessels. Images with blood vessel metrics above a threshold $t_v$ were classified as gradeable while images with metrics below $t_v$ were classified as ungradeable. Usher's algorithm was evaluated using 800 images from 400 patients, comparing results to the opinion of three clinicians. An average inter-grader agreement was calculated using the average agreement between the system and the individual clinicians. Usher's algorithm reportedly achieved 100% sensitivity and 94% specificity in detecting patients with at least one ungradeable image.

Fleming *et al.* [7] similarly base the quality grading on blood vessel segmentation; however, they limit the analysis to a macula-centred square region 3.5DD in width, based on the position determined by their fovea-location algorithm. If the fovea location algorithm has a poor correlation coefficient (indicating uncertain location) they search for the least-dense

3.5DD region within a 4.5DD area around the estimated fovea centre.

The guidelines for human grading of retinal image quality are carefully designed to reflect the problem domain, and it follows that automated techniques should be based upon these criteria. Usher *et al* [5,6] come close to achieving this but overlook several issues. First, blood vessels can be visible without being in focus so that blurred vessels may add to the blood vessel pixel count. Second, no emphasis is given to macula vessels, which are the most diagnostic. Consequently, as macula vessel only count for a small percentage of the total vascular network, images with limited or no macula vessels could still be classified as gradeable. Third, displacement from the defined position is not addressed. Fleming *et al.* [7] address two of these issues, but their metric does not address blurred vessels other than by possible failure to segment.

## III. Method

The image assessment algorithm proposed below is based upon the UK National Screening Committee guidelines and addresses all the issues discussed above. The approach calculates the contrast and quantity of visible blood vessels within 1DD of the fovea, and measures the contrast between the foveal core region and the background retina.

The optic nerve head is first located and measured using the algorithm in [8]. Then, the fovea is located by selecting the maximum correlation point with a 40×40 Gaussian filter with standard deviation $\sigma = 22$ pixels [9]. The vascular map is segmented, skeletonized and morphologically ``cleaned up'' using the tramline algorithm [10]. The quality analysis is then conducted in a circular region of interest of radius two optic disk diameters around the estimated foveal centre.

The appearance of small blood vessels within 1DD of the fovea is the primary indicator of fundal image quality. Three aspects of macula vessels contribute to the quality measure – distance from fovea, and the clarity as represented by the proxies of contrast and quantity. For each vascular segment, $i \in \{1, S\}$, the number of pixels in the segment, $\eta_i$, the average distance of the pixels from the fovea centre, $\gamma_i$, and the contrast with the local background retina, $\alpha_i$, are calculated. The contrast, $\alpha_i$, is measured by subtracting the average intensity of the segment centreline pixels from the average intensity of segment boundary pixels, where the boundary pixels are calculated by dilating the segment centerline using a $6 \times 4$ and a $4 \times 2$ structuring element and taking the logical difference (based on an assumption that macula vessels are less than seven pixels in diameter). An overall vascular metric is then defined as:

$$\nu = \sum_i \frac{\eta_i \alpha_i}{\gamma_i} \qquad (1)$$

which sums the product of segment contrasts and lengths, penalized by distance from the centre of the fovea.

The contrast of the foveal region to the retinal background in the macula gives a secondary indicator of fundal image quality. This is estimated by calculating the contrast, $\kappa$, between the average intensity of a circular region with radius $r = 10$ (half that of an average fovea radius) with the average intensity of a doughnut-shaped region with inner and outer radii $r_1 = 30$ and $r_2 = 60$, both centred on the estimated fovea centre. The overall image quality measure $\mu$ is given by:

$$\mu = \upsilon\kappa \qquad (2)$$

The UK national screening guidelines for image quality divide images into three quality categories: *achievable*, *minimum* and *ungradeable*. We used a more detailed five category scale, which has proven more useful for integration with automated diagnostic algorithms. Images with small blood vessels visible around the fovea and with good foveal contrast with the background macula area are graded as 1, images with similar vascular detail but with reduced contrast are graded 2. Images that only include macula periphery blood vessels are graded 3 if there is good foveal contrast, 4 otherwise. If no vessels were visible, the image is graded as 5. Grades 1-2 correspond to achievable, grade 3-4 to minimum and grade 5 to ungradeable images. The algorithm determines the category by binning $\mu$; the bin thresholds were empirically determined to minimize the category error using a reference set of 100 images categorized by an ophthalmologist. Table I lists the resulting category ranges.

TABLE I
RANGE OF $\mu$ FOR FIVE QUALITY CATEGORIES

| Category | Minimum | Maximum |
|---|---|---|
| 1 | 331 | - |
| 2 | 101 | 330 |
| 3 | 36 | 100 |
| 4 | 6 | 35 |
| 5 | 0 | 5 |

## IV. Evaluation

The performance of the macula image quality assessment algorithm is benchmarked against two alternate techniques. A reference dataset was constructed, consisting of 200 760×570 fundus images randomly drawn from a diabetic retinopathy screening dataset. An ophthalmologist graded these on the 1–5 scale described above.

The presented algorithm was compared with Lalonde's [2] template intensity histogram and Usher's [5] vascular metric. The benchmark algorithms were implemented so that a comparison could be made using a single data set, rather than comparing to the results presented in [2,5] on different datasets. In [5] the vascular metric consisted of the sum of all pixels contained within the blood vessel network. In this study it was found more reliable to morphologically thin the segmented blood vessels to a centreline and sum the vascular centreline pixels. This reduces metric variability due to blood vessel width and treats all blood vessels with equal importance, whereas Usher's algorithm is heavily influenced by the wider major temporal retinal vascular arcades, so that macula vessels having little influence on the overall vascular metric. It is against this modified algorithm that the macula model is evaluated.

Lalonde stated that images could be crudely categorized into three groups: "good", "fair" and "bad". However, in testing Lalonde's algorithm on the 200 screening images we found a three-way split impractical due to heavy overlapping of the measure between categories, and hence graded as either gradeable or ungradeable.

This application is sensitivity biased as it is critical to correctly identify 100% (or as close to that as possible) of ungradeable images, thus avoiding any potential subsequent misdiagnosis due to poor image quality. For the first evaluation, we therefore evaluated the algorithms for determination of ungradeable (category 5) images. We used Receiver Operating Characteristic (ROC) curves to select the lowest thresholds that achieve 100% sensitivity for each algorithm (the thresholds used were 5, 4586.31 and 11419 for the proposed model, Usher and Lalonde's respectively). Table II shows the results of the presented and alternate algorithms.

TABLE II
COMPARATIVE PERFORMANCE OF ALGORITHMS

| Statistic | Proposed | Usher | Lalonde |
|---|---|---|---|
| Sensitivity | 100 | 100 | 100 |
| Specificity | 93 | 87 | 19.5 |
| Accuracy | 94 | 88 | 28 |

Usher [5] reported a specificity of 94% at sensitivity 100%; however, in our evaluation the specificity was 87%. In our evaluation Lalonde's algorithm achieves a poor specificity of 19.5% at 100% sensitivity, but results dramatically improve by allowing one false negative classification, giving 95% sensitivity and 81% specificity. Our algorithm achieved 93% specificity, almost halving the number of false positives compared to [5].

TABLE III
QUALITY GRADING PERFORMANCE BY GRADE (SCALE 1-5)

| Grade | Correct | Incorrect | % Correct |
|---|---|---|---|
| 1 | 25 | 1 | 96 |
| 2 | 33 | 0 | 100 |
| 3 | 20 | 3 | 85 |
| 4 | 15 | 4 | 73 |
| 5 | 7 | 0 | 100 |

For automated diagnostic algorithms it may be more important to have a finer level of gradation of image quality. In our second experiment we evaluated our algorithm's ability to grade the images on the 1-5 scale using 100 of the clinically assessed images (the other 100 being used to set the thresholds). Table III shows that 91% of automated image quality assessments matched the clinician; the remaining 9% are all within one grade of the clinician's assessment. It is worth noting the system detected 100% of clinically ungradeable images (grade 5). The high categorisation accuracy of this approach means that automated image quality assessment can exclude not only clinically ungradeable images, but also borderline cases (category 4) leaving only the highest quality images for automated classification.

## V. CONCLUSION

This paper has presented an algorithm for automated quality rating of retinal images. The algorithm is based on UK National Screening Committee guidelines for image quality assessment, and achieves a good level of agreement with clinical assessments. It identifies image quality with higher reliability than benchmark methods, and can provide a finer level of gradation. Further work is required to include optic disc clarity and image alignment, as described in the National Screening Committee's (NSC) [1]) image quality guidelines.

## ACKNOWLEDGMENT

This work was supported under Diabetes UK grant No. BDA:RD00/0002033.

## REFERENCES

[1] P.H. Scanlon, R. Malhotra, G. Thomas, C. Foy, J.N. Kirkpatrick, N. Lewis-Barned, B. Harney, and S.J. Aldington, "The effectiveness of screening for diabetic retinopathy by digital imaging photography and technician ophthalmoscopy," Diabetes UK. *Diabetic Medicine*, vol. 20, pp. 467 474, 2003.

[2] S.C. Lee and Y. Wang, "Automatic retinal image quality assesment and enhancements," in *Proc SPIE* Vol. 3661 Med Imaging:Img Processing, 1999, pp. 1581–90.

[3] M. Lalonde, M. Beaulieu, and L. Gagnon, "Automated visual quality assessment in optical fundus images," in *Vision Interface*, pp. 259–264, 2001.

[4] M. Niemeijer, M.D. Abràmoff and B. van Ginneken. "Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening," vol. 10 no. 6, pp. 888-898, 2006.

[5] D.B. Usher, M. Himaga, M.J. Dumskyj, J.F. Boyce, A. Sabate-Cequier, and Williamson T.H., "Automated assessment of digital fundus image quality using detected vessel area," *Proceedings of medical image understanding and analysis the university of sheffield*, pp. 81–84, 2003.

[6] M. Himaga, D. Usher, and J.F. Boyce, "Retinal blood vessel extraction by using multiresolution matched filtering and directional region growing segmentation," *Proc. IAPR Workshop on Machine Vision Applications*, Nara, Japan, pp. 244–247, 2002.

[7] A. Fleming, S. Philip et al., "Automated assessment of diabetic retinal image quality based on clarity and field definition", *Investigative Opthalmology and Visual Sciences* vol. 47 no. 3, pp. 1120–1125, 2006.

[8] J.A. Lowell, A. Hunter, D. Steel, B. Ryder, E. Fletcher. "Optic Nerve Head Segmentation" *IEEE Transactions on Medical Imaging*, vol. 23 no. 2, pp. 256-264, Feb 2004.

[9] C. Sinthanayothin, J.A. Boyce, H.L. Cook, and T.H. Williamson, "Automated localization of the optic disc, fovea, and retinal blood vessels from digital colour fundus images," *Br J. Ophthalmol*., vol. 83, pp. 902–910, 1999.

[10] A. Hunter, J. Lowell and D. Steel, "Tram-line filtering for retinal vessel segmentation." *In: Proceedings of the 3rd European Medical and Biological Engineering Conference*, EMBEC05, Prague, Czech Republic, 2005.