Automatic visual quality assessment in optical fundus images

Marc Lalonde†, Langis Gagnon† and Marie-Carole Boucher‡
†Centre de recherche informatique de Montréal
550 Sherbrooke W., Suite 100, Montréal, Qc, H3A 1B9
{mlalonde, lgagnon}@crim.ca
‡Département d'ophtalmologie, Hôpital Maisonneuve-Rosemont
5689 Boul Rosemont, Montréal, Qc, H1T 2H1

Abstract

This paper presents a method for automatically assessing the quality of retinal images. It is based on the idea that images of good quality possess some common features that should help define a "quality" model. The proposed features are computed from the histogram of the edge magnitude distribution in the image as well as the local histograms of pixel gray-scale values. Histogram matching functions are proposed and experiments tend to show that these features help discriminate between good, fair and bad images.

1 Introduction

Automatic image quality assessment (IQA) is a major issue in medical imaging, but it is most often viewed as a signal processing or degradation/restoration problem whose relevance is most acute when e.g. evaluating lossy compression methods for image storage or transmission ([2], [1]). A much less frequently studied issue is having to decide whether a particular image is suitable for diagnosis purposes. This latter topic is especially pertinent in the context of telemedicine operations where a remote operator is in charge of taking images that will eventually be analyzed by a physician. In this context, the operator may not be familiar with the criteria guiding the assessment of images and some form of automated support should be provided.

Our focus is more specifically on quality assessment of low-resolution (640x480) ophthalmic images captured by a high-quality, 3-CCD color camera. Although noise is kept to minimal levels[1], a number of factors may degrade the image:

- the patient's eye might be blinking during image acquisition, so eyelashes or, worse, the eyelid may be imaged in the process;
- the patient might move his head during the examination, which would result in bad image acquisition (image is out of focus, part of the iris is imaged, image appears darker because the retina is no longer properly illuminated, etc.).

In addition, image quality happens to be dependent upon the type of diagnosis being made. In other words, a particular image with dark regions might be considered of good quality for detecting glaucoma but of bad quality for detecting diabetic retinopathy.

2 Previous work

Previous work in the area of automatic IQA of ophthalmic images is quite limited. To our knowledge, besides some references to IQA in research reports (e.g. [7]), only Lee and Wang ([4]) have dedicated a paper to this topic. However their work is guided by signal processing concerns since they suggest that their method can be used to objectively quantify the gain or loss of quality following some restoration / enhancement of an image. Their approach can be summarized as follows. They define a template intensity histogram whose parameters were derived from analysis of 20 images with excellent quality (from a set of 360 images) and whose base (its width or spread) can be interpreted as a measure of contrast in the image (the contrast is viewed as a key feature); the quality of a target image is assessed by convoluting its histogram with the template histogram and by computing a quality index Q. The index Q has a value between 0 and 1, with Q≈0 meaning that the image is of very poor quality. However, our own analysis of about 40 retinal images of varying quality tends

[1] These cameras typically have a >60dB signal-to-noise ratio (SNR).

to show that the connection between image quality and histogram similarity is not that strong. In other words, we found some bad quality images whose histogram resembled the template histogram and we also found good quality images with markedly different histograms. Let us precise that our notion of quality is defined with respect to one's capacity of using the image for the diagnosis of diabetic retinopathy, and in that context, an improved method could be developed.

3 Description of the proposed approach

The approach described in this paper is generally similar to that of Lee and Wang (1999) in the sense that a model of what is a good image is defined using a set of images of excellent quality. However, the model itself is different and stems from observations on the characteristics of good and bad images. Two criteria have been retained which focus on:

1. the distribution of the edge magnitudes in the image, and
2. the 'local' distribution of the pixel intensity, as opposed to the global histogram of Lee and Wang.

In the next section, we will review the criteria and show how they can be combined together to help characterize the quality of a target image.

3.1 Edge magnitude distribution

It has been noted that the distribution (histogram) of the edge magnitudes in a good ophthalmic image has a shape that is similar to a Rayleigh distribution but with a gentler drop as the intensity increases, whereas a bad image sees its edge distribution fall rapidly. This is illustrated in Figure 1 which shows a good (top) and a bad (bottom) image along with their edge distributions.

Although a remarkably bad image has been selected in the example, a quick examination of additional bad images shows that the histogram spread is also limited or quite different from that of known good images. So a natural criteria would be a measurement of the difference between the edge magnitude histogram of a target image and a 'typical' edge magnitude histogram. The typical histogram was constructed with the edge maps of a set of good images, and the difference is evaluated using an equation similar to the χ^2 statistic for comparing two binned data sets ([6]):

$$d_{edge}(T,R) = \sum_i \frac{(R_i - T_i)^2}{R_i + T_i}, \quad \forall\, i \,|\, R_i + T_i \neq 0$$

where R is the reference histogram and T is the edge histogram of the target image. Note that bins which were empty for both R and T simultaneously were removed from the computation in order to avoid divisions by zero. Figure 2 shows the typical edge distribution that serves as a model.

This criteria might be loosely interpreted as a measure of focus in the sense that well-focused images have clear and sharp borders, which translates into high magnitude edges.

3.2 Intensity distribution

It is clear that the distribution of gray-scale values also plays a role in one's assessment of the image quality. A good image should not possess too many dark or white pixels and the mean gray-scale value should be similar to what one would consider a perfect image. We adopt Lee and Wang's approach of defining an ideal intensity histogram from examples drawn from a set of good images but instead of building a global measure of similarity between the intensity distribution of the whole target image and an ideal distribution model, we advocate a region-based approach that is broken down into the following steps:

1. an ideal gray-scale image I_{Mean} is created with the aggregation of all images from a 'perfect' set;
2. the target image I_{Target} is then segmented into uniform regions by a standard segmentation algorithm (e.g. a histogram-splitting algorithm adapted from [5]); smaller regions are discarded;
3. for each segmented region in I_{Target}, a histogram is constructed; obviously, the histogram features very few non-empty bins since the region is believed to be of uniform grey level by virtue of the segmentation process. A histogram is also built from the same region in I_{Mean}. This implies a prior alignment of the two images, which is performed by translating one image with respect to the other so that the two circular apertures overlap.
4. the dissimilarity measure between I_{Target} and I_{Mean} is computed as follows:

$$d_{intensity}(I_{Target}, I_{Mean}) = \sum_{i=1}^{Nb\,regions} Size_i \cdot W(HR_i^{(I_{Target})}, HR_i^{(I_{Mean})})$$

where $HR_i^{(I_{Target})}$ is the histogram of the i^{th} segmented region, drawn from the target image, $HR_i^{(I_{Mean})}$ is the histogram of the same region drawn

from the mean image and $Size_i$ weighs the contribution of each pair of histograms according to the size of the region being considered (small regions have a smaller impact). Finally, the matching function W measures the dissimilarity between the two histograms and is defined by:

$$W(h_1, h_2) = \left[\frac{\mu_{h_1} - \mu_{h_2}}{\min(\mu_{h_1}, \mu_{h_2})}\right]^2$$

The function W is designed to measure the 'alignment' of the histograms by computing the difference between their means. This difference is divided by the smallest mean in order to penalize misalignment of low-mean histograms (this choice is explained by the fact that regions of low intensity have a serious impact on image quality since structures and lesions may be less visible). In this case, the histogram itself is not used; in fact, one could compute $d_{intensity}$ without histogramming the regions. The use of histograms in this context results from multiple tests having been performed by the authors in order to devise an appropriate matching function W.

3.3 Decision

Deciding whether the image I_{Target} is of good or bad quality may bear some resemblance to a classification problem where the two computed measures earlier, d_{edge} and $d_{intensity}$, act as features. In this light, the next step is to evaluate whether a decision border exists that can separate the classes "Good", "Fair" and "Bad". All the images of our dataset were first classified by an ophthalmologist, then the 'features' were extracted from each image. Figure 3 illustrates the separability of the classes. Since only 40 images were available for this study, random sub-sampling was performed with a small subset of the 'good' images used for building the edge and intensity models. Of the 40 images of the dataset, for each experiment eight good images were drawn at random and used to build the models; the rest were used for testing. The points in Fig. 3 represent the outcome of assessing the quality of all images in the dataset, the location of a particular point being set according to the values of $d_{intensity}$ and d_{edge} for the corresponding image and its symbol (+, x or o) being chosen according to the ophthalmologist's assessment. The ellipses illustrate the scatter in each 'class' for eight experiments, with their principal axes determined according to the standard deviation of the variables. The ellipses are plotted for illustrative purposes; they do not suggest that the class densities follow a known (e.g. Gaussian) distribution, so one should look at the plot with caution. However, if straight decision borders were to be established based upon the plot (i.e. placed so as to separate the classes, in the spirit of Bayes decision theory), one would get the following confusion matrix:

Expert ↓System→	Good	Fair	Bad
Good	16	2	1
Fair	3	8	2
Bad	0	1	7

Obviously, many more data would be required to optimize the type and placement of the borders (they may not be straight lines); nevertheless, the plot suggests that the 'features' $d_{intensity}$ and d_{edge} might be appropriate for automatic discrimination between good and bad images. In some cases the system may 'decide' that an image is good or bad while the human observer considers it fair (or vice versa) but we view these errors to be acceptable, unlike cases where both judges have totally different opinions. This latter case happens once in our experiments, and the image involved is shown in Figure 4. The subject's small pupil makes the image much darker at the periphery of the retina, which in turn distorts the edge distribution histogram. The strong deviation from the models leads the system to label the image as 'bad', yet the expert maintains that some form of diagnosis is possible in the central portion of the image, so it should be viewed as 'good'.

4 Discussion

Automating an image quality assessment (IQA) process is a challenging task. Typically, a human observer will carry out the evaluation with specific criteria in mind that are strongly related to the image contents. For example, he/she might rate an image 'bad' because the region around the macula is too dark. Yet mimicking this "behavior" would require prior knowledge of the location of structures such as the macula. From the viewpoint that IQA would be the first action performed on the image following acquisition, this knowledge would be unavailable and an IQA module would base its decision upon more "down-to-earth" features such as overall contrast, focus measure, etc. From another viewpoint, one could see IQA as a "recurring" process triggered whenever some new knowledge is extracted from the image.

In any case, experiments tend to show that one reliable feature is provided by the edge magnitude distribution. Our parallel with a focus measure seems to be appropriate, especially in light of the work done in depth from focus (e.g. [3]) which proposes and uses focus measures to automatically change the focus of the vision system in order to estimate depth. Indeed, a sim-

ple experiment of defocusing an image using a Gaussian or a mean filter showed that d_{edge} increases as the blurring gets more pronounced. As for the intensity criterion, the results are interesting but a bit less conclusive. Such a criterion may be defined in many different ways, not only in terms of distance between histograms (ours is based on misalignment of histogram means) but also in terms of spatial location. For example, misalignment of histograms drawn from specific regions (e.g. those that are close to some structures, if such information were to be available) might be more penalized.

Of course, a better evaluation of the proposed features would require many more images, and ideally a comparison with many human observers so as to quantify the correlation between the 'response' of each feature and the judgment of the specialists.

5 Conclusions

In this paper, a first set of criteria has been proposed for automatic image quality assessment in the context of retinal image analysis. Two criteria were retained: 1) a measure of match between the edge magnitude distribution of the image and a model derived from a set of good images, and 2) a measure of match between intensity histograms of some regions of the image and those from a model image resulting from averaging of the same set of images. Tests with a set of 40 images show that both criteria help discriminate between good and bad images. Additional tests with a larger set of images and comparison against the opinion of many ophthalmologists are needed for a better evaluation of the performance of the approach.

References

[1] P.C. Cosman, R.M. Gray, and R.A. Olshen. Evaluating quality of compressed medical images: Snr, subjective rating, and diagnostic accuracy. *Proc. of the IEEE*, 82(6):919–932, June 1994.

[2] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C Bovik. Image quality assessment based on a degradation model. *IEEE Trans. on Image Processing*, 9(4):636–650, April 2000.

[3] T. Darell and K. Wohn. Depth from focus using a pyramid architecture. *Pattern Recognition Letters*, 11:787–796, December 1990.

[4] S.C. Lee and Y. Wang. Automatic retinal image quality assessment and enhancement. In *Proc. SPIE Conf. on Image Processing*, pages 1581–1590, Feb. 1999.

[5] R. Ohlander, K. Price, and D.R. Reddy. Picture segmentation using a recursive region splitting method. *Computer Graphics and Image Processing*, 8:313–333, 1978.

[6] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical recipes in C : the art of scientific programming.* Cambridge University Press, 1992.

[7] OPHTEL Project. Image processing within OPHTEL. `http://www-ophtel.gsf.de/OPHTEL/ima_proc/ophtel_img_proc.html`, March 2000.

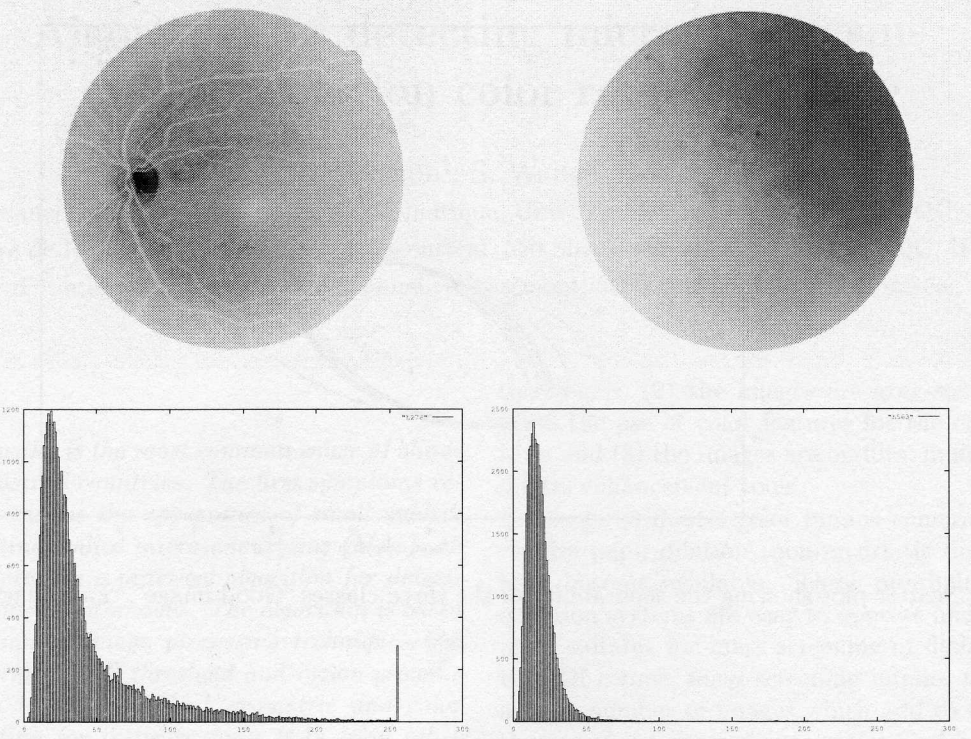

Figure 1: Examples of a good and a bad image, with their corresponding edge distributions. Note that the gray-scale images have been inverted for better visibility.

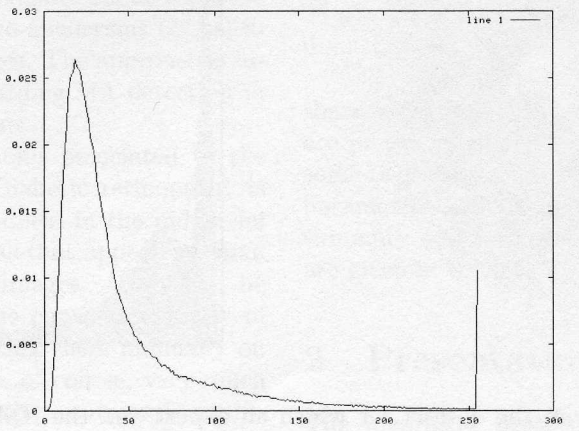

Figure 2: Edge distribution model.

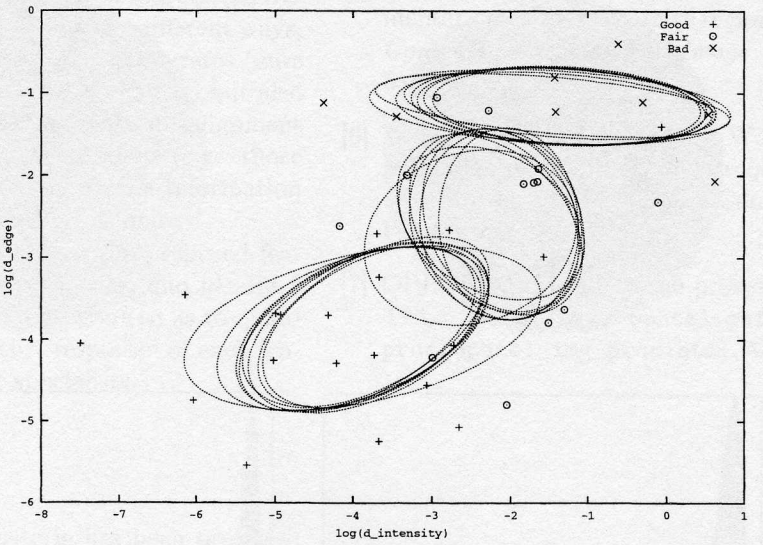

Figure 3: Scatter plot showing the separability of the three classes "Good image", "Fair image" and "Bad image".

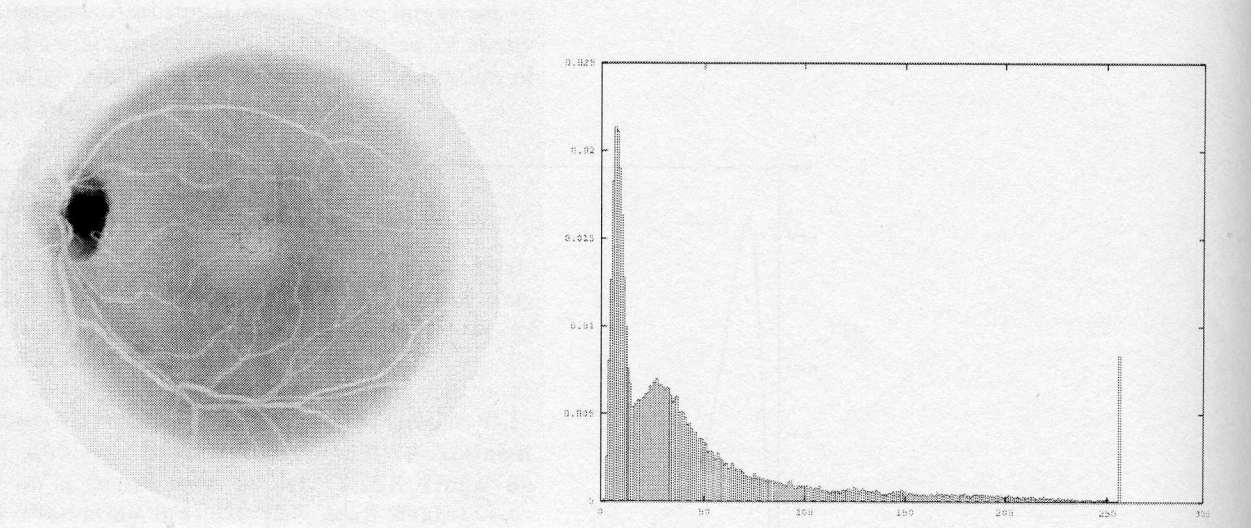

Figure 4: (Left) Image from the dataset for which an expert and the IQA method gave conflicting responses. Note that the image is inverted for better visibility. (Right) Edge distribution of the image.