Retinal Image Quality Classification Using Neurobiological Models Of The Human Visual System

Dwarikanath Mahapatra
dwarim@au1.ibm.com.

IBM Research Melbourne, Australia

Abstract. Retinal image quality assessment (IQA) algorithms use different hand crafted features without considering the important role of the human visual system (HVS). We solve the IQA problem using the principles behind the working of the HVS. Unsupervised information from local saliency maps and supervised information from trained convolutional neural networks (CNNs) are combined to make a final decision on image quality. A novel algorithm is proposed that calculates saliency values for every image pixel at multiple scales to capture global and local image information. This extracts generalized image information in an unsupervised manner while CNNs provide a principled approach to feature learning without the need to define hand-crafted features. The individual classification decisions are fused by weighting them according to their confidence scores. Experimental results on real datasets demonstrate the superior performance of our proposed algorithm over competing methods.

1 Introduction

Image quality assessment (IQA) of retinal fundus images is an important step in screening systems for diseases like diabetic retinopathy (DR). Automated analysis requires retinal images to be of a minimum quality that would facilitate feature extraction. Figures 1 (a)-(b) shows examples of ungradable images that hamper reliable feature extraction.

Reliable factors for IQA identified by the Atherosclerotic Risk in Communities (ARIC) [1] study are grouped into two major categories: generic image quality parameters (e.g. contrast, clarity, etc) and structural quality parameters (such as visibility of the optic disc and macula). Methods using generic image information include histogram matching [10] and distribution of edge magnitudes [9]. Despite low computational complexity, they do not always capture diversity of conditions affecting image quality. Other IQA methods use retinal landmarks like the vasculature [15] and multi scale filter banks [12]. They require anatomical landmark segmentation which is complex and error prone, especially for poor quality images. Paulus et al. [14] combined generic and structural image features but rely heavily on accurate landmark segmentation.

X. Chen, M. K. Garvin, J. Liu, E. Trucco, Y. Xu (Eds.): OMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, Iowa Research Online, pp. 97–104, 2016. Available from: http://ir.uiowa.edu/omia/2016_Proceedings/2016/

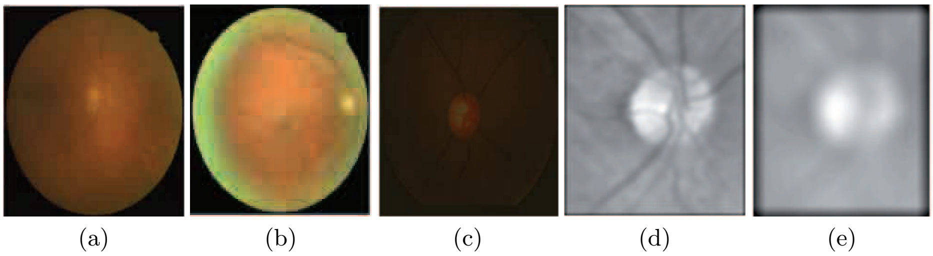


Fig. 1. (a)-(c) Examples of ungradable images; (d)-(e) example outputs of convolving learned kernels with original image.

Humans rely on the human visual system (HVS) to identify poor quality images. IQA is subjective as it depends on a user's perception of good quality. Current approaches to IQA use hand crafted features which do not generalize well to new datasets. Neither do they leverage the functioning of the HVS to improve IQA. This necessitates solving the problem using computational principles behind the working of the HVS, thus minimizing subjectivity and bias of existing algorithms. We propose a method for retinal IQA that uses computational algorithms imitating the working of the HVS. Our objective is achieved through the following novelties: 1) We propose a novel 'local saliency map' that calculates saliency values for every image pixel across different scales, and captures local and global image information that is relevant for IQA; 2) we leverage learned supervised information from convolutional neural networks (CNNs) thus avoiding hand crafted features. We combine supervised (trained CNNs) and unsupervised (local saliency maps) models using Random forest (RF) classifiers and the associated confidence scores, and demonstrate their superior performance over competing methods.

2 Methods

2.1 Saliency Model

The original 8 bit color images are intensity normalized to $[0 - 1]$ and resized to 512×512 pixels.Saliency defines the degree to which a particular region is different from its neighbors with respect to image features. The original model by Itti-Koch [7] gives a global saliency map highlighting attractive regions in the image. Visual input is first decomposed into a set of multiscale feature maps and different spatial locations compete for saliency within each map. These feature maps are combined to form a final saliency map that highlights the most salient regions in an image. The limitation of state of the art saliency algorithms [7, 5, 6] is they highlight a single region that is most salient and pixels outside the salient region have no importance. We propose a 'local' saliency map method that calculates the saliency value of each pixel by incorporating principles of neurobiology into the algorithm. Since image quality assessment should incorporate both local and global features, our saliency maps incorporate them by taking multiple scales (neighborhoods) of each pixel.

The resized color image is converted to gray scale intensity, and texture and curvature maps are obtained from this grayscale image. Multiscale saliency maps are generated from these 3 feature maps. According to neurobiological studies, the response function of cortical cells is Gaussian [2], i.e., further away a point, less is its influence on the central pixel. Thus, to calculate a pixel's uniqueness from its surroundings a sum of weighted difference of feature values is calculated,

$$D_F(s) = \sum_i \exp\left(-\left\|s - s_i\right\|\right) \left|F(s) - F(s_i)\right|, \tag{1}$$

where D_F indicates the difference map for feature F; s_i is the ith pixel in the $N \times N$ neighborhood of pixel s; $\|s - s_i\|$ denotes the Euclidean distance between s and s_i. $F(s_i)$ denotes the feature value at pixel s_i. This gives a saliency value for each pixel. We use different values of N (5×5, 11×11, 19×19, 31×31 and 41×41) to get saliency maps for intensity, texture and curvature at varying scales for capturing local and global information.

Figure 2 shows the 'global' saliency maps generated by different methods and the local saliency maps obtained by our method at different scales for an original 'gradable image'. Figure 2 also shows the corresponding saliency maps for an ungradable image. A comparative study of the maps highlights the following points: 1) our local saliency maps provide more discriminative information than the global saliency maps which only highlights the optic disc region as the most salient region. 2) the local saliency maps are able to capture different levels of local and global information by varying the operation scale. 3) local saliency maps are more effective than global saliency maps in discriminating between gradable and ungradable images. These set of results justify our proposed local saliency maps instead of using the conventional saliency maps.

5 different scales for the 3 saliency maps gives a total of 15 saliency maps. Each map is divided into non-overlapping 64×64 blocks, giving a total of 64 blocks. The mean pixel value of each block is calculated to give 64 feature values for one map. The total number of features from the 15 maps is $(64 \times 15 =)960$ which is the feature vector obtained from saliency maps.

2.2 CNN Architecture

Figure 3 (a) shows the architecture of our proposed network. The input patches are of size 512×512 which are put through 5 layers of convolution and max pooling operations. The first convolution layer $C1$ takes as input the 512×512 patch and convolves it with 10 11×11 kernels to return a 512×512 output. The patches are symmetrically padded to ensure that the convolution output is of the same size as the input. The details of the number of kernels and their sizes are depicted in Figure 3 (a). $C1$ 10/11/512 denotes that layer $C1$ has 10 kernels of size 11×11 and outputs a 512×512 patch. The max pooling layer is

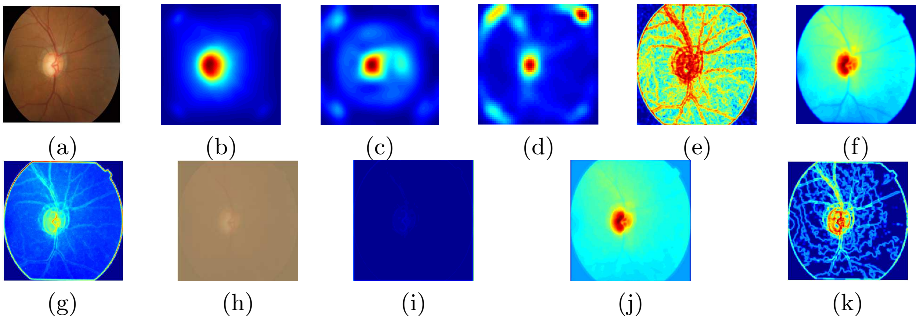


Fig. 2. (a) original gradable image; (b)-(d) global saliency maps obtained using [7, 5, 6]; (e)-(g) local saliency maps from our method at different scales; (h) ungradable image; (i)-(k) corresponding local saliency maps by our method.

2×2 with $512 \rightarrow 256$ indicating the input 512×512 patch is downsampled to 256×256 using max pooling. Following $C5$ (the fifth convolution layer) we have 3 fully connected layers of $4000, 2000, 1000$ nodes followed by a soft-max classifier that outputs the class label as either gradable or ungradable. We refer to this architecture as CNN_5, i.e., a CNN with 5 convolution layers. The soft-max layer has a logistic regression that calculates the probability of each class as,

$$P\left(y = i|\mathbf{W},\mathbf{b}\right) = \frac{\exp^{W_i x + b_i}}{\sum_{j=1}^{M} \exp^{W_i x + b_i}}, \tag{2}$$

where x is the output of the second fully connected layer, W_i and b_i are the weights and biases of the i^{th} neuron in this layer. The class with maximum probability is the predicted class.

Instead of traditional sigmoid or tanh neurons, we use Rectified Linear Units (ReLUs) [11] in the different layers since recent research [8] has demonstrated the speedup in training compared to using tanh units. An ReLU has an output of $f(x) = max(0; x)$ where x denotes the input. ReLUs enable the training to complete several times faster and are not sensitive to the scale of input.

2.3 Training the CNN

We use negative log-likelihood as the loss function and perform Stochastic Gradient Descent (SGD) using dropout where the neuron outputs are masked out with probability of 0.5, and at test time their outputs are halved. Dropout alleviates overfitting by introducing random noise to training samples and boosts the performance of large networks. Since applying dropout to all layers significantly increases the training time, we only apply dropout at the second fully connected layer, i.e., half of the outputs of the second fully connected layer are randomly masked out in training, and in testing the weights of the logistic regression layer are divided by 2, which is equivalent to halving the outputs of the second fully connected layer. Figure 3 (b) shows example learned filters from the final convolutional layer.

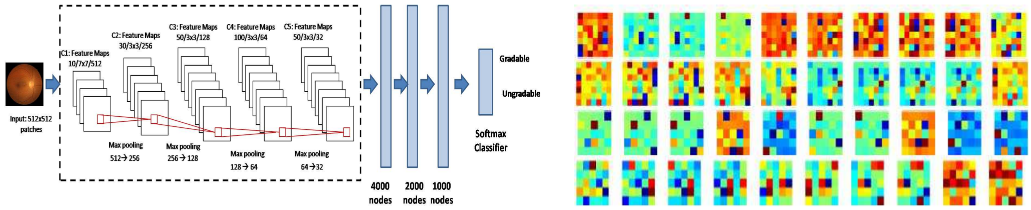
2.4 Image Quality Classification

The feature vector from saliency maps (f_1) and the 1000 dimensional feature vector from the last fully connected layer of the CNN (f_2) are used to train two different Random forest (RF) classifiers [3] (denoted as RF_1 and RF_2). RF_1 and RF_2 are both trained on the image labels. A given test image is resized to 512×512 and put through the process of saliency map generation and CNN classification to output two class labels (0/1 for ungradable/gradable images) alongwith probabilty scores. Thus for every test image we have two probability values (for gradable/ungradable) each from RF_1 and RF_2.

The probability values act as confidence scores for each classifier which is used to calculate the final label (C) as,

$$C = \frac{w_{1,1} + w_{2,1}}{2}, \tag{3}$$

where $w_{1,1}$ is the confidence score (probability) of RF_1 predicting class 1 and $w_{2,1}$ is the confidence score of RF_2 predicting class 1. If $C > 0.5$ then the final prediction is class 1 (gradable), else the image is deemed ungradable. The advantage of this approach is the combination of supervised (RF_1) and unsupervised (RF_2) image features. Note that the CNN has a probabilistic classifier which also outputs probability score, and in principle, there is no need to train a separate RF_2 classifier. However we do that for the sake of continuity with saliency features, although our experiemntal results show there is no significant performance difference if we use the soft max classifier instead of RF_2.



(a) (b)

Fig. 3. (a) Illustration of the CNN architecture used for our proposed method; (b) Examples of learned filters from the final convolutional layer.

3 Experiments and Results

3.1 Dataset Description

We use a dataset acquired from a DR screening initiative. The dataset ($D1$) has 9653 ungradable retinal images and 11347 gradable images. All of the images are non-mydriatic, have a 45° FOV and a resolution of 2812 × 2442 pixels. All the images have been graded by human graders, thus confirming their gradability labels.

Data Augmentation: Since the dataset size is not large enough to train a robust CNN we apply data augmentation by image translation and horizontal reflections to increase the number of images. All image intensities were normalized between $[0,1]$ and then resized to 512 × 512 pixels. These resized images were subject to different operations like horizontal and vertical flipping, rotation, translation and contrast changes. Using these operations the size of our dataset was increased 50 times, which is large enough to avoid overfitting. Henceforth further references to any dataset refers to the augmented version. $400,000$ images from each class of the augmented dataset are used to train the CNN. We use a $5-$ fold cross validation approach where the training data consists of 80% of the total images. We ensure that the augmented versions of an image are either in the training or test set.

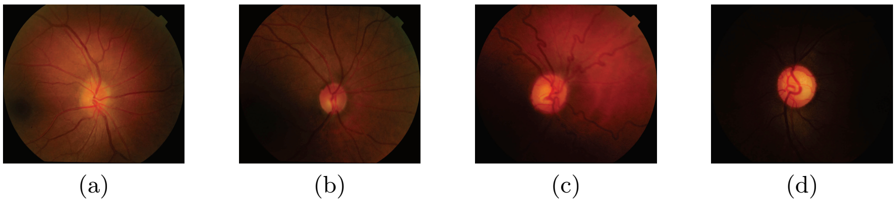
Classification Results: Results of our method (denoted RF_{1+2}) are compared with the following methods : RF_{All} where the feature vectors f_1, f_2 are concatenated to train a RF; SVM_{All} - support vector machines using f_1, f_2 with linear kernels for classification; $RF_1 + SM$ - weighted combination of outputs of RF_1 and the CNN softmax classifier for predicting the gradability. To perform a comparative study, we have tested the methods proposed in [4], [14] and [12]. We re-implement these algorithms by closely following the details given in the respective works.

Table 1 shows Sensitivity (Sen), Specificity (Spe) and Accuracy (Acc) obtained using $5-$fold cross validation. We obtain high sensitivity (correctly identified gradable images), specificity (correctly identified ungradable images) and accuracy values which outperforms current state-of-the-art methods for our dataset. These values are higher than those reported in the original works of [4, 12] and significantly better than [14]. A significant achievement of our method is that it has been tested on a much larger dataset than previous works. The dataset covers a wide range of images acquired under different conditions and provides a much stricter evaluation of different algorithms. The $p-$values from a paired $t-$test with the results of RF_{1+2} show the significant improvement brought about by neurobiological models of the HVS. $p < 0.05$ indicates that the two sets of results being compared are statistically significant.

The kernel sizes in different layers were decided based on extensive experiments where effect of different kernel sizes on final accuracy was studied. We have used progressively smaller kernel sizes in different layers in order to capture information at different scales. The first layer has 7 × 7 kernels which captures more global information. From the second convolution layer onwards the kernel

	RF_{1+2}	RF_{All}	SVM_{All}	Paulus [14]	Dias [4]	Niemeijer [12]	RF_1+ SM	RF_1	SM
Sen	98.2	95.4	95.1	94	96.1	96.7	97.9	92.2	93.4
Spe	97.8	94.6	94.2	90.1	95.4	96.0	97.8	91.8	92.4
Acc	97.9	94.7	94.5	91.4	95.6	96.2	97.9	91.9	92.8
$p-$	-	0.0012	0.0018	0.00009	0.0017	0.0024	0.56	0.0001	0.0001

Table 1. Sensitivity (Sen), Specificity (Spe), Accuracy (Acc) and $p-$values for different methods compared to CNN.



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

Fig. 4. (a)-(b) Gradable images which were classified incorrectly by other algorithms except RF_{1+2} as ungradable; (c)-(d) ungradable images classified as gradable by all algorithms except RF_{1+2}.

size is fixed at 3×3. Since the image dimensions are progressively halved in each layer, the fixed size kernels capture a mix of local and global information.

Results from Table 1 also show the advantages of fusing the decisions of RF_1, RF_2 instead of concatenating them in a single feature vector (RF_{All}, SVM_{All}). It also shows that the softmax classifier ($RF_1 + SM$) performs as good as the RF classifiers. Figures 4 (a)-(b) show examples of gradable images which were correctly classified by RF_{1+2} as gradable but incorrectly classified by [4], [14] and [12]. This was probably due to uneven illumination and intensity saturation at some parts. Figures 4 (c)-(d) show the opposite case where ungradable images were classified as gradable by [4], [14] and [12] but not by RF_{1+2}. A major factor behind the superior performance of our method is the use of CNNs and saliency maps which: 1) are not dependent on hand crafted features, and hence generalize well; 2) combination of supervised and unsupervised features.

Computation time: The average computation time for classifying a test image is 8.2 seconds with our method using non-optimized MATLAB code on a Intel Core 2.3 GHz $i5$ CPU running Windows 7 with 8 GB RAM. Although not real time, classification time is small enough to make a quick decision about repeat scans. The CNN architecture was trained using the MATLAB Deep Learning Toolbox [13]. The average training time for $400,000$ patches from each class is 22 hours. Feature extraction from saliency maps and its classification takes 3.2 seconds while feature extraction from CNNs and classification takes 4.7 seconds with a further 0.3 seconds for fusing the two decisions.

4 Conclusion

We have proposed a novel method to determine image quality of acquired retinal scans by combining unsupervised information from visual saliency maps and supervised information from trained CNNs. Our key contribution is the use of computational models of HVS for IQA, and an algorithm for local saliency map computation. We also extract additional information from trained CNNs. Combining these two sources of information leads to high sensitivity and specificity of our method which outperforms other approaches. The low computation time is an added benefit for a quick assessment of image quality in settings which require a quick decision to determine whether the patients would need a repeat scan.

References

1. The atherosclerosis risk in communities (ARIC) study: design and objectives. The ARIC investigators. am j epidemiol. 1989 apr; 129(4), 687-702.
2. E. Goldstein, Sensation and perception, Thomson Wadsworth, 2007
3. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
4. Dias, J., Oliveira, C., Cruz, L.: Retinal image quality assessment using generic image quality indicators. Information Fusion 19, 73–90 (2014)
5. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems (NIPS). pp. 545–552 (2006)
6. Hou, X., Harel, J., Koch, C.: Image signature: Highlighting sparse salient regions. IEEE Trans. Pattern Anal. Mach. Intell. 34(1), 194–201 (2012)
7. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20(11), 1254–1259 (1998)
8. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 1106–1114 (2012)
9. Lalonde, M., Gagnon, L., Boucher, M.: Automatic visual quality assessment in optical fundus images. In: Proc. Vision Interface. pp. 259 – 264 (2001)
10. Lee, S., Wang, Y.: Automatic retinal image quality assessment and enhancement. In: Proc. SPIE Medical Imaging. pp. 1581–1590 (1999)
11. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: International Conference on Machine Learning. pp. 807–814 (2010)
12. Niemeijer, M., Abramoff, M., van Ginneken, B.: Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. Med. Imag. Anal. 10(6), 888–898 (2006)
13. Palm, R.B.: Prediction as a candidate for learning deep hierarchical models of data (2012)
14. Paulus, J., Meier, J., Bock, R., Hornegger, J., Michelson, G.: Automated quality assessment of retinal fundus photos. Intl. J. Comp. Assisted Radiology and Surgery . 10(6), 888–898 (2006)
15. Usher, D., Himaga, M., Dumskyj, M.: Automated assessment of digital fundus image quality using detected vessel area. In: Proc. Medical Image Understanding and Analysis. pp. 81–84 (2003)