IMPACT OF IMAGE PREPROCESSING ON EYE FUNDUS SEGMENTATION

Arnaud Robert

Department of Computer and Software Engineering, Polytechnique Montreal, Canada

ABSTRACT

While most advances and innovations are focused on improving the architecture of CNN, comparatively little work has been done on studying the impact of image treatment prior to entering the network. Thus, in this paper, we present a comparative study of the impact from different types of image preprocessing on eye fundus image segmentation. Our network consists in the classic U-net architecture to which minor modifications were made. The structure of the eye can be very complex and a vast array of biomarkers can be part of the fundus image, depending on the patient's health status. For the purposes of this study, we will focus on a single kind of structure: vessels (arteries and veins combined). Our model is trained on a hundred images from the MESSIDOR database, then tested on 97 images from the same database. The results show that most filters do not improve on performance by a large margin. Only the median filter yields consistent improvement on performance. Finally, the study indicates that there is a need to define a good framework (metrics to use, tests to conduct) through which relevant comparisons can be made.

***Index Terms*—** image preprocessing, segmentation, CNN, U-net, deep learning

1. INTRODUCTION

Fundus imaging is defined by Abràmoff et al. [1] as:

> "The process whereby a 2-D representation of the 3-D retinal semi-transparent tissues projected onto the imaging plane is obtained using reflected light."

The images obtained through this process can be used to detect retinal diseases such as diabetic retinopathy and age-related macular degeneration, through the detection of physical symptoms inside the retina. They can also help clinicians determine risk factors for systemic diseases such as cardiovascular disorders.
Efficient methods for analysing fundus images in the past decade have involved the use of deep convolutional neural networks (CNN). The most successful ones have been using networks based on the U-net architecture [2] and achieved remarkable results such as 98,02% AUC score for vessel segmentation [3] on the DRIVE dataset. We observe among the literature that the usage of image preprocessing is variant, ranging from using a single channel filtered with CLAHE and Gamma correction in Qiangguo et al. (2018) [3] to not using preprocessing at all. This situation results from the fact that the impact of preprocessing on neural networks is not well understood in the community. Image preprocessing can have many effects such as data normalisation, noise reduction or feature enhancement, all of which are used in order to achieve better performance. Is it commonly believed in the community that, ideally, a neural network would be able to extract relevant features from an unprocessed image, thus replicating the effect of a preprocessing algorithm, and indeed succeeding in learning the appropriate preprocessing by itself. In practice, however, applying preprocessing steps does have an effect on the end result, which is why many authors tend to use it. One could also argue that some $basic$ filters, such as the median filter, can be unnecessary hard to learn for the network.

This paper investigates the effects of image preprocessing on the task of eye fundus image segmentation. To do so, we train several networks using different preprocessing algorithms and compare their results.

2. METHOD

2.1. Network architecture

Our network is a modified U-net [2] consisting of four learning blocks in each path with a bottleneck. One block consists of two 3x3 convolutions with padding (to avoid cropping the image trough the network), a scaled exponential linear unit (SeLU) as an activation function, and a 2x2 max pooling operation (in the encoder) or a 3x3 upsampling convolution (in the decoder). The final block in the decoder ends with a 3x3 padded convolution.

2.2. Preprocessing techniques

Preprocessing is used to improve the quality and characteristics of the image in order to achieve superior performance during training. Namely, it can be used to reduce variance, remove noise and augment contrast. There are infinitely many ways to treat the image in order to achieve these goals and hence we will limit ourselves to a total of 5 different techniques.

2.2.1. *CLAHE*

Initially developed for medical imaging, the Contrast Limited Adaptive Histogram Equalization [4] is a popular contrast enhancement method. It subdivides the image in a grid of contextual regions for which the optimal contrast is calculated, taking into account that there exists homogeneous regions, such as the image background, for which the contrast enhancement has to be limited, thus avoiding to make background noise visible.

2.2.2. *Individual color channel*

By taking a look at each individual colour channel of an eye fundus image, one can realise not all of them convey the same amount of information. It can be argued that the blue channel has too little contrast, and the red channel is simply too saturated. The green channel however, seems to have the best contrast and carries relatively little noise. We will investigate the correctness of these statements and their hypothetical impact on segmentation by training a model for each individual channel of the image.

2.2.3. *Median filter*

Using a sliding 3x3 kernel over the image, each pixel is replaced by the median value across the kernel, thus smoothing and removing noise in the image. The resulting image is then subtracted from the original one, which effectively reduces the variance among all images.

2.2.4. *Morphoprocess*

A downside of using the median filter as described previously is that it deteriorates the edges of the image. This deformation can later interfere with the network training and be mistaken for a vessel. We propose a preprocessing method for tackling this concern. We first perform a dilatation on the edges of the eye fundus image. The median filter is then calculated on the result of this operation. Finally, it is subtracted from the image through a mask that encapsulates only the area of the initial image. Hence, this filter effectively avoids taking background pixels into account when performing the median filter and instead relies on a dilatation of the border pixels to compute the median values at the border.

2.2.5. *Filters combination*

A last filter is made by simply combining two existing ones, namely the green channel only filter and the morphoprocess one. Note that the order of application does not matter since the morphoprocess is applied channel-wise.

2.3. Training

The cross entropy loss function is used during training. We also use the Adam optimizer algorithm with an initial learning rate of 0.0001 and weight decay equal to 0.1.

3. EXPERIMENTAL RESULTS

3.1. Datasets

For training and testing we used a random 197 images subset of the Messidor-2 [5] dataset which consists of 1748 images from diabetic retinopathy screenings. These images were annotated by clinicians providing separate ground truth images for vessels, red lesion and bright lesions. 100 images were devoted to training while the 97 remaining ones were used for testing. Data augmentation in the form of random rotation through a 180 degrees angle is being used as well.

3.2. Results

We chose to use the Cohen's kappa coefficient, recall and precision-recall AUC at the pixel level as metrics for our comparison. These metrics were chosen due to their relative robustness to the presence of class imbalance. The kappa coefficient compares expected and observed agreement levels while taking into account the fact that said agreement may be the product of chance [6]. The mean results over the 97 Messidor test images can be seen in figure 1.

Preprocessing	P-R AUC	Recall	Kappa
raw	0.739369	0.772186	0.791883
red	0.435040	0.480664	0.465158
green	0.711226	0.818880	0.801651
blue	0.088427	0.556574	0.092347
clahe	0.704312	0.803177	0.785470
median	0.749722	0.826898	0.809515
morph	0.719023	0.761205	0.773316
morph+green	0.705102	0.766283	0.771667

Fig. 1. Mean results on Messidor test set.

As expected, the red and blue channel only treatments yield poor results. The other preprocessing methods seem to have some effect, the magnitude and significance of which have to be evaluated. To do so, we conduct a statistical test on each metric, namely the Mann-Whitney U test, which is a non parametric alternative to the Student t-test. This test indicates whether or not it is reasonable to affirm that the median from two independent sets are similar. The null hypothesis (H_0: the median from both sets are similar) is rejected if the resulting p-value is inferior to a given significance level. Here we require the p-value to be inferior to 0.01 to reject H_0. Tables in figure in figure 2 show the resulting p-value for each possible combination of preprocessing. One table is produced for each of our three metrics. The color of a cell indicates whether the null hypothesis is rejected, green corresponding to rejection at a 1% level. Each black colored cell indicates, for each preprocessing algorithm, the mean value of the corresponding metric.

Using the raw image as a base for comparison, we observe that, among all preprocessing treatments, the median filter appears to be the only one to always improve the performance, as measured by the mean metric, and to do significantly so. For the Recall metric, the green filter appears to significantly improve the performance as well. Interesting to note is that the impact of the green and median filters on this metric can not be deemed different with significant certainty.

Oddly enough, we note that when it comes to the Recall metric and the blue filter, p-values behave strangely. For example, the impact of the morphoprocess and the blue filter can not be considered different according to the p-value, even though the mean metric scores, 56% and 76% respectively, are clearly different. This may suggest that a median centered test such as the one we are using may not be valid under certain extreme cases.

4. CONCLUSIONS

This paper presented a comparative study of the impact of various preprocessing techniques on eye fundus vessel segmentation, using a U-net architecture. The results indicate that using a median filter yields the most consistent improvement for the task, and that its effect is significant. Other techniques do not have a considerable positive effect on all our metrics simultaneously. Some other metric may be more relevant for evaluating the performance and could be added to gain more insight on the consistency of the filters. One can also argue about the extent to which comparing median values, which is what our statistical test does, is relevant.

The main challenge faced in the study was defining a relevant way to compare the results. There is a wide variety of metrics that can be employed, and comparing the preprocessing techniques through a same metric can also be done in numerous ways, using different tests. Future work should define a good framework through which results would be compared.

5. REFERENCES

[1] K. Lee W. L. W. Alward Y. H. Kwon et M. D. Abramof L. Tang, M. K. Garvin, “Robust multiscale stereo matching from fundus images with radiometric differences.,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2245–2258, 11 2011.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” *CoRR*, vol. abs/1505.04597, 2015.

[3] Qiangguo Jin, Zhao-Peng Meng, Tuan D. Pham, Qi Chen, Leyi Wei, and Ran Su, “Dunet: A deformable network for retinal vessel segmentation,” *CoRR*, vol. abs/1811.01206, 2018.

[4] Karel Zuiderveld, “Graphics gems iv,” chapter Contrast Limited Adaptive Histogram Equalization, pp. 474–485. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[5] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, and Jean-Claude Klein, “Feedback on a publicly distributed database: the messidor database,” *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, Aug. 2014.

[6] Arie Ben-David, “Comparison of classification accuracy using cohen’s weighted kappa,” *Expert Systems with Applications*, vol. 34, pp. 825–832, 02 2008.

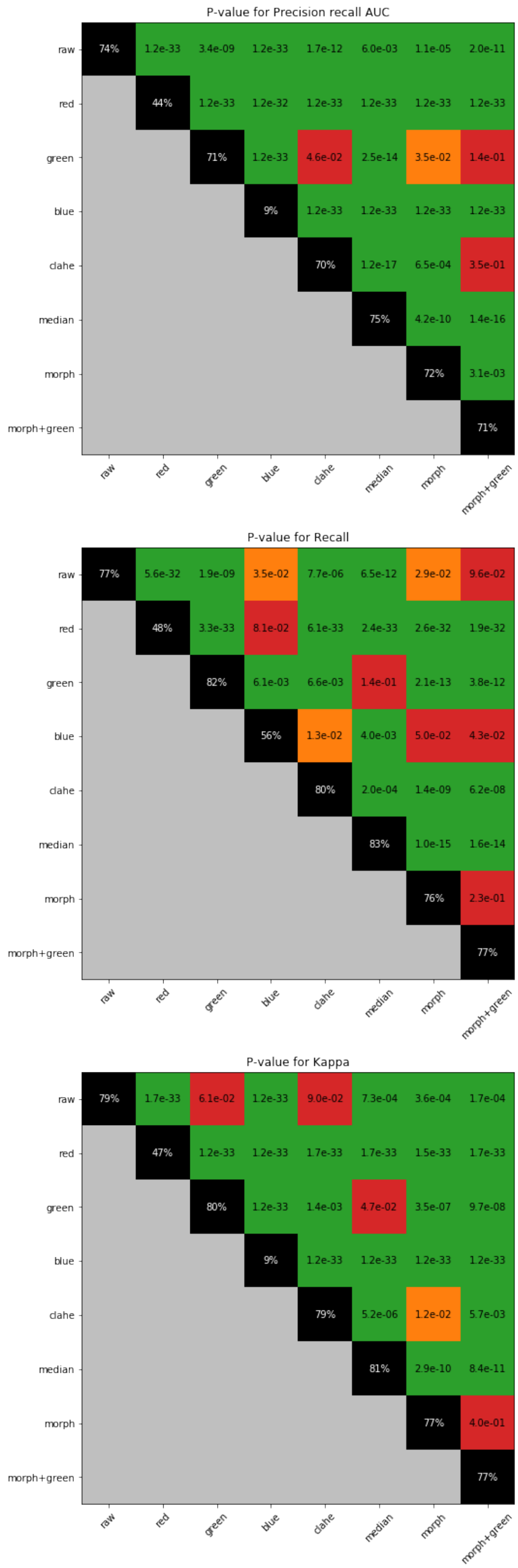

Fig. 2. Cross treatments P-values for each metric.