Journal of Engineering and Applied Sciences 11 (11): 2378-2382, 2016
ISSN: 1816-949X

Benchmarked Pterygium Images for Human and Machine Graders

[1]Mohd Zulfaezal Che Azemin, [1]Norfazrina Abdul Gaffur, [1]Mohd Radzi Hilmi,
[2]Mohd Izzuddin Mohd Tamrin and [3]Khairidzan Mohd Kamal
[1]Kulliyyah of Allied Health Sciences, Pahang, Malaysia
[2]Kulliyyah of ICT, Selangor, Malaysia
[3]Kulliyyah of Medicine, International Islamic University Malaysia, Kuantan, Malaysia

Abstract: In the absence of ground truth, scores from many graders are required to obtain good representation of a clinical grading. The internet enables quick feedback from the experts at the comfort of their home or office. In this study, we demonstrated the use of online form as a tool to get quick feedback from clinicians on clinical grading of pterygium images with various severities. The scores were analyzed using quartile analysis and the median was used to construct the benchmark scores for the images. This dataset was tested on assessing human grader and was later fitted with neural network to measure the performance of the machine learning algorithm.

Key words: Machine learning, benchmarked dataset, ground truth, quick feedback, Malaysia

INTRODUCTION

Pterygium is a type of benign growth that occurs in the eye region that is not obstructed by the eyelids when the eye is opened. It can cause blindness when it remains untreated for a period of time. Pterygium grading is commonly developed based on the image characteristics of the fibrovascular tissues. Tan *et al*. (1997) for example, devised a grading system based on the translucency of the tissue. Grade 1 is given when episcleral vessels under the pterygium body are not covered and clearly seen and Grade 3 is associated with increased in fleshiness and when the episcleral vessels are completely covered. Another grading system of pterygium is based on the encroachment of the pterygium body (Mahar and Manzar, 2013). The level of severity is graded based on the extension of the tissues towards pupils, the opening located in the center of the eye which allows light to be focused on retina.

Redness in the eye is related to blood vessel dilation in the conjunctiva and sclera regions. It is another means of assessing severity of the damage in the tissue (Murphy *et al*., 2007). This approach is currently underexplored for the assessment of the pterygium tissue. Grading redness is commonly done subjectively by clinicians using a set of reference images as a guide (Schulze *et al*., 2011). While this method is considered repeatable, it must be done by trained graders. Automating this process will result in a highly consistent grading, regardless of the experience of the graders (Wu *et al*., 2015).

The first step of automating a subjective grading is to define the ground truth that we consider as the benchmark of the score. The judgment of more than one experienced graders is commonly employed when there is absence of device that is able to accurately measure a clinical observation (Trucco *et al*., 2013). This study demonstrated the use of benchmarked pterygium images evaluated by clinical experts to assess the performance of human grader and machine learning algorithm.

MATERIALS AND METHODS

Figure 1 shows the overview of the methodology employed in this study. A total of 68 pterygium images were taken using slit-lamp biomicroscopy at a controlled exposure. A subset of 30 images was carefully selected to represent different type of severity.

Figure 2 illustrates an online form developed using Google form to collect the scores from other clinicians. Three benchmarked images (Fig. 3) were graded by a pterygium specialist (Grade 1 = the least red, Grade 2 = moderate, Grade 3 = the most red). To increase the objectivity of the grading, the region of interest was restricted to apex to limbus of the pterygium.

Collection of benchmarked data: Google Form was selected as a platform for its ease of use, free, and the data can be exported to major spreadsheet formats for further processing. The form contains the pterygium images and dropdown box with 5-point scale points (i.e., Grade 1,

Corresponding Author: Mohd Zulfaezal Che Azemin, Kulliyyah of Allied Health Sciences, Pahang, Malaysia

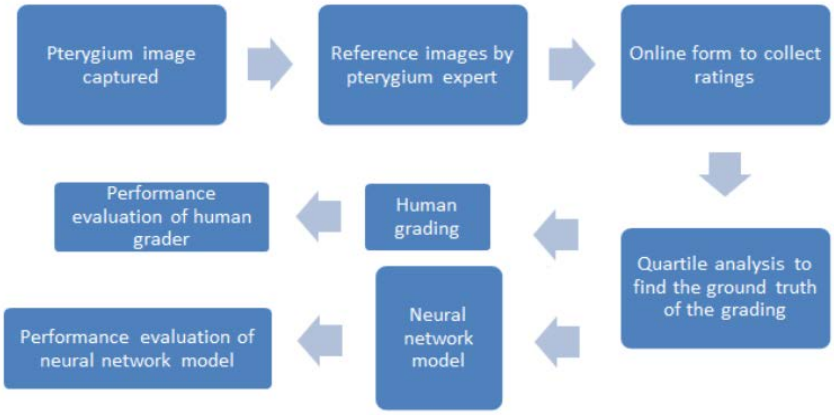

Fig. 1: Block diagram of the overview of the research

Fig. 2: Reference images for grading purpose

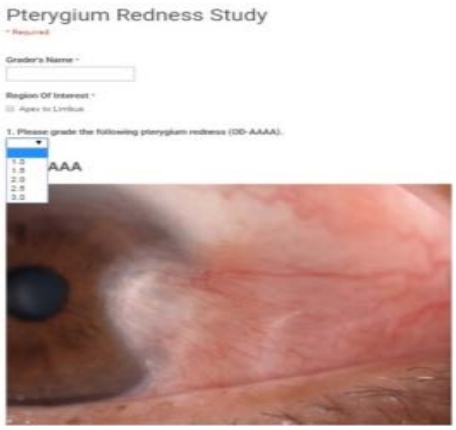

Fig. 3: Online form to collect data from graders

1.5, 2.0, 2.5 and 3). The clinicians must have at least 2 year clinical experience and familiar with other ocular redness grading (e.g., Efron or CCLRU). To complete the survey an average grader requires approximately 5 min. The images graded by clinicians as the least (i.e., Grade 1) and the most red (i.e., Grade 3) by 5-point scale including the intermediate grades (e.g., Grade 1.5) based on reference images rated by pterygium specialist. Since, the reference and graded images are viewed on the same screen, the illumination and contrast of both set of images are comparable.

Human grader: To demonstrate the use of this dataset to assess the performance of human graders, we have evaluated the performance of two human graders (Grader W and R) against the ground truth. Grader W has less than 3 years of experience in the clinical grading and does not primarily involve in pterygium cases, as opposed to Grader R with >5 years of experience in pterygium and clinical grading.

Machine grader: The performance for machine learning algorithm was also evaluated using this benchmarked data. Artificial Neural Network (ANN) has been chosen to model the grading of the pterygium redness. ANN has shown to be promising in learning non-linear complex data. To avoid over-fitting, supervised machine learning commonly requires a training set. The number of the images was increased to 68 images using scores from human expert which gives good agreement with the benchmarked images as the ground truth. A total of 34 images were used for the training phase and 34 more images were used in the testing phase. The training and testing data were independent to each other.

As a first step to model the neural network, a total of 5 features were extracted from the images. The features were computed based on textural analysis (e.g., homogeneity and contrast) on multiple color space (e.g., RGB, YUV, HSI), extracted from three different channels (e.g., Red, Green and Blue from RGB color space). The features were identified from a total of 211 features using Maximal Relevance Minimal Redundancy (MRMR) algorithm (Peng *et al*., 2005). The MRMR algorithm filters penalizes the relevant of a feature by looking at its redundancy relative to the other features.

Three features were selected from images transformed from RGB to YUV color space. The YUV color space

commonly employed between source of the image (e.g., scanner) and image generator (e.g. display monitor). It is designed with human perception in mind particularly the careful choosing of the chrominance components. The conversion from RGB was done using Eq. 1:

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.14713 & -0.2886 & 0.436 \\ 0.615 & -0.51499 & -0.1001 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

The fourth feature was selected from images converted from RGB to HSI color space. HSI was initially used to strike a balance between image segmentation performance and processing power. The I component represents the intensity average of the RGB components. The S component is the saturation value as defined by Eq. 2:

$$S = [1 - \min(R,G,B)/I] \quad (2)$$

The H component in Eq. 3 is characterized by the human colors perception (i.e., yellow, red, green and blue) (Getreuer, 2015). It is modeled based on the following equation:

$$H = \tan^{-1}(\beta, \alpha) \quad (3)$$

Where:

α = 1/2 (2R-G-B)

β = $\sqrt{3}/2(G-B)$

The fifth selected feature comes from CIE XYZ color space which was developed from two experiments conducted with 17 observers. The color space attempts to mimic the average person's experience of color sensation. The XYZ values are derived from the spectral power distribution of these average observers.

Following the color space conversions, the final features were calculated based on the pattern from Gray-Level Co-occurrence Matrix (GLCM) (Haralick *et al.*, 1973). The GLCM was quantified using the statistical analyses; correlation, homogeneity, energy and contrast. The features were fed into a regularized feed-forward artificial neural network trained by Levenberg-Marquardt backpropagation. The neural network has the following parameters:

- Size of hidden layer = 10
- Maximum number of epochs for training = 300
- Performance measure = Mean-squared error
- Performance goal = 1×10^{-5}

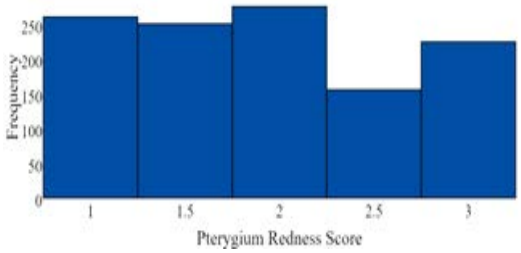

Fig. 4: The distribution of the grading data from 39 graders

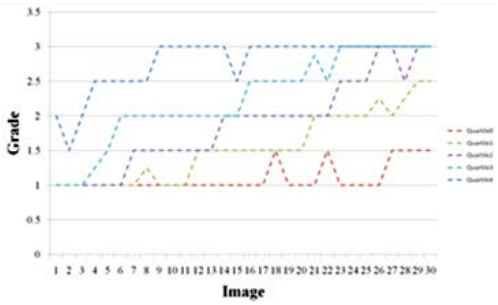

Fig. 5: Quartile analysis of the graded images

To find the compromise between high bias (under-fit) and high variance (over-fit), the value for the regularization parameter was varied from 0-1 with an increment of 0.05. The regularization parameter that gives the lowest minimum-squared error was selected as the optimum value.

RESULTS AND DISCUSSION

A total of 39 graders responded to the survey, coming from optometry and ophthalmology backgrounds. The average of the redness score is 1.93 with standard deviation of 0.71. Figure 4 summarizes the respondents' data as a histogram. It shows an evenly distributed data, suggesting the image dataset covers redness severity from least to most severe cases.

In this study, we adopted the definition of ground truth similar to the previous study (Fieguth and Simpson, 2002). Median has been used as the ground truth for the bulbar redness. Median is more favored over average because it is less susceptible from outliers. Figure 5 shows the quartile analysis of the grading data. For visualization convenience, the images were sorted based on the average score. Quartile 0 and 4 represent the extreme scores and Quartile 2 is the median of the data.

Performance evaluation of human grader: Figure 6 shows the Bland-Altman analysis of the grading by Grader W

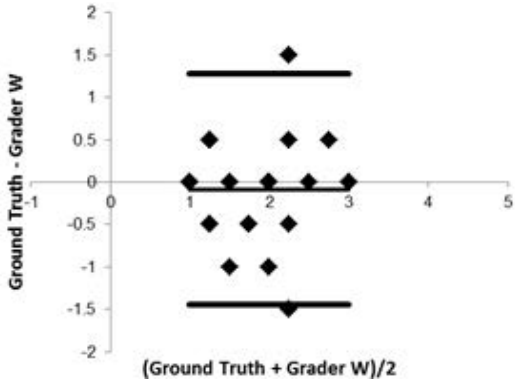

Fig. 6: Bland-Altman plot of grader W against the ground trust

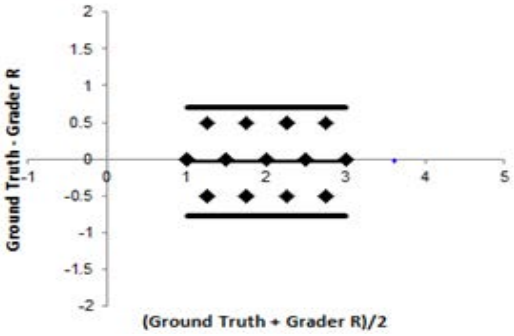

Fig. 7: Bland-Altman plot of grader R against the ground truth

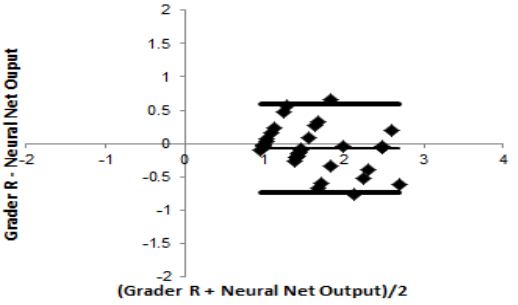

Fig. 8: Bland-Altman plot of output of neural network against the scores graded by grader R

when compared with the ground truth. It is observed that the average difference of 0.083 with limits of agreement (LOA) of -1.45 and 1.28, it shows moderate intra-class correlation (ICC=0.655) with the ground truth.

Grader with more clinical experience exhibits closer grading score with the ground truth with an excellent intra-class correlation (ICC = 0.921). Bland-Altman plot (Fig. 7) shows smaller average difference of -0.03 and LOA of -0.77 and 0.70.

Performance evaluation of machine grader: In supervised learning, more data are required to split the data into training and testing sets. Since, grader R demonstrates an excellent agreement with the ground truth, we had used the grader's scores as the labels for the performance evaluation of the machine grader. The human expert (grader R) graded 38 additional images in addition to the original 30 images, giving the total image of 68. Split validation (50% training vs 50% testing) was used to assess the performance of the neural network model.

The artificial neural network exhibits an excellent agreement with grader R (ICC = 0.898). The Bland-Altman plot in Fig. 8 shows average difference of 0.07 with the LOA of -0.74 and 0.59. This implies the ANN model is able to mimic the grading of the human expert. Figure 8 Bland-Altman plot of output of neural network against the scores graded by grader R.

CONCLUSION

Researchers have demonstrated the possibility of developing clinical image dataset with its respective grading based on data extracted from an online form. These benchmarked images were shown to be useful in assessing the performance of human and machine learning algorithm. The performance of a newly developed algorithm can also be tested using dataset in the future (Azemin *et al*., 2014).

ACKNOWLEDGEMENTS

This research was supported by the Ministry of Higher Education of Malaysia under the Fundamental Research Grant Scheme with identification n umber FRGS14-138-0379.

REFERENCES

Azemin, C., M. Zulfaezal and M. Hilmi, 2014. Supervised pterygium fibrovascular redness grading using generalized regression neural network. N. Trends Software Methodologies Tools Tech., 2014: 650-656.

Fieguth, P. and T. Simpson, 2002. Automated measurement of bulbar redness. Invest. Ophthalmol. Visual Sci., 43: 340-347.

Haralick, R.M., K. Shanmugam and I. Dinstein, 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern., SMC-3: 610-621.

Mahar, P.S. and N. Manzar, 2013. Pterygium recurrence related to its size and corneal involvement. J. Col. Physicians Surg. Pak, 23: 120-123.

Murphy, P.J., J.S.C. Lau, M.M.L. Sim and R.L. Woods, 2007. How red is a white eye? Clinical grading of normal conjunctival hyperaemia. Eye, 21: 633-638.

Peng, H., F. Long and C. Ding, 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell., 27: 1226-1238.

Schulze, M.M., N. Hutchings and T.L. Simpson, 2011. Grading bulbar redness using cross-calibrated clinical grading scales. Invest. Ophthalmol. Visual Sci., 52: 5812-5817.

Tan, D.T., S.P. Chee, K.B. Dear and A.S. Lim, 1997. Effect of pterygium morphology on pterygium recurrence in a controlled trial comparing conjunctival autografting with bare sclera excision. Arch. Ophthalmol., 115: 1235-1240.

Trucco, E., A. Ruggeri, T. Karnowski, L. Giancardo and E. Chaum *et al*., 2013. Validating retinal fundus image analysis algorithms: Issues and a proposalvalidating retinal fundus image analysis algorithms. Invest. Ophthalmol. Visual Sci., 54: 3546-3559.

Wu, S., J. Hong, L. Tian, X. Cui and X. Sun *et al*., 2015. Assessment of bulbar redness with a newly developed keratograph. Optometry Vision Sci., 92: 892-899.