Automated Image Quality Assessment for Fundus Images in Retinopathy of Prematurity

Aaron S. Coyner[1,2], Ryan Swan[1,2], Jayashree Kalpathy-Cramer[3], Sang Jin Kim[1,4], J. Peter Campbell[1], Karyn E. Jonas[5], Susan Ostmo[1], R.V. Paul Chan[6], Michael F. Chiang[1]

[1]Ophthalmology, [2]Medical Informatics, Oregon Health & Science University, Portland, OR United States.
[3]MGH/Harvard Medical School, Boston, MA, United States.
[4]Ophthalmology, Sungkyunkwan University School of Medicine, Seoul, Korea.
[5]University of Illinois at Chicago, Chicago, IL, United States.
[6]Ophthalmology, Illinois Eye and Ear Infirmary, Chicago, IL, United States.

SUPPORTED BY RPB Research to Prevent Blindness

PURPOSE

To implement an algorithm for automatically assessing the quality of fundus images acquired from preterm infants with a possible diagnosis of Retinopathy of Prematurity (ROP).

INTRODUCTION

- Accurate image-based ophthalmic diagnosis depends upon the clarity of fundus images.
- Emerging techniques, such as telemedicine and computer-based image analysis, perform suboptimally when provided with low-quality images.
- Low-quality images necessitate the need for additional imaging sessions, resulting in undue stress on patients, as well as wasted time and resources for physicians and researchers.
- Implementation of an algorithm capable of assessing imaging quality could negate the abovementioned consequences.
- Said algorithm could be implemented at the point of care, or used as a preprocessing step before images are used in telemedicine or computer-based image analysis.

METHODS

IMAGE EVALUATION

1. Thirty wide-angle images were acquired from premature infants.
2. Images were scaled and converted to grayscale.
3. Each image was broken into 16-by-16 pixel blocks.
4. Each block was assessed for quality using brightness and contrast metrics.
5. Acceptable blocks received a score of 1, unacceptable neighborhoods received a score of 0.
6. The sum of the acceptable quality blocks in the image was divided by the total number of blocks in the image to provide an overall image quality score (range 0 to 1).

ANALYSIS

1. Six expert image graders were selected to evaluate the quality of the same set of 30 images.
2. Each grader's image ranking, from worst to best, was determined using an Elo rating system.
3. A consensus rank was determined using the ranks from the six individual graders.
4. A Spearman correlation value was calculated between the consensus rank of the images, the algorithm rank of the images, and each individual expert grader's rank of the images.

RESULTS

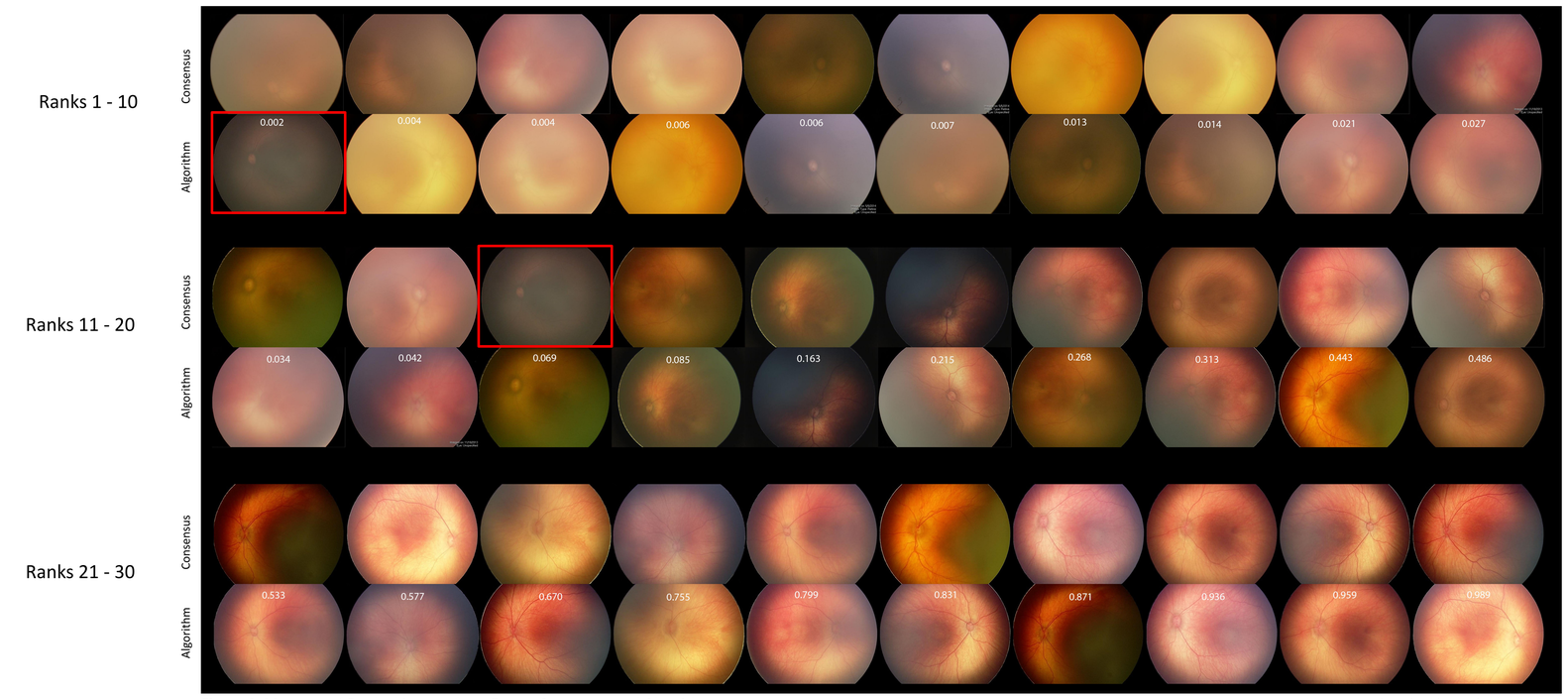

Figure 1: Montage of Ranked Images from Expert Consensus and Algorithm
Images, ranked 1 (worst quality) through 30 (best quality), are displayed in three rows. Each row shows the image in its corresponding rank for the expert consensus rank (top sub-row) and the algorithm rank (bottom sub-row). Qualitatively, the algorithm is able to separate images of lower quality from images of higher quality. Red box illustrates a discrepancy between the algorithm rank and the consensus rank.

Table 1: Correlation Matrix of Expert Grader Rank and Algorithm Rank
Spearman's rank test correlation matrix of expert consensus rank, algorithm rank, and ranks from the six individual experts.

	Consensus	Algorithm	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6
Consensus	1.00	0.86	0.98	0.94	0.94	0.97	0.96	0.96
Algorithm	0.86	1.00	0.82	0.81	0.82	0.86	0.87	0.86
Expert 1	0.98	0.82	1.00	0.90	0.89	0.94	0.93	0.93
Expert 2	0.94	0.81	0.90	1.00	0.94	0.96	0.92	0.95
Expert 3	0.94	0.82	0.89	0.94	1.00	0.91	0.91	0.92
Expert 4	0.97	0.86	0.94	0.96	0.91	1.00	0.94	0.97
Expert 5	0.96	0.87	0.93	0.92	0.91	0.94	1.00	0.97
Expert 6	0.96	0.86	0.93	0.95	0.92	0.97	0.97	1.00

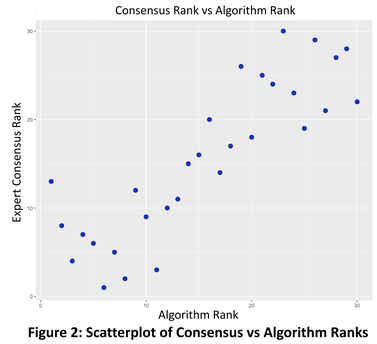

Figure 2: Scatterplot of Consensus vs Algorithm Ranks
Each point represents an image and its corresponding expert consensus rank (y-axis) and algorithm rank (x-axis).

CONCLUSIONS

- Expert grader ranks were highly correlated with one another (correlation coefficient [CC] 0.89-0.94).
- Individual expert grader ranks were highly correlated with the consensus rank (CC 0.94 – 0.98).
- The algorithm rank was moderately correlated with the consensus rank (CC 0.86) and individual expert ranks (CC 0.81- 0.87).
- Discrepancies between the algorithm and consensus ranks need to be further investigated.
- Qualitatively, the algorithm adequately separates low- from high-quality images.
- Future modifications will be assessed using the established consensus rank as a reliable training set.

ACKNOWLEDGEMENTS

Supported by NLM Training Grant 4T15LM007088-25, NIH Grant P30, and unrestricted departmental funding from Research to Prevent Blindness.

REFERENCES

1. Chiang MF, Wang L, Busuioc M, Du YE, Chan P, Kane SA, Lee TC, Weissgold DJ, Berrocal AM, Coki O, Flynn JT, Starren J. Telemedical Retinopathy of Prematurity Diagnosis Accuracy, Reliability, and Image Quality. Arch Ophthalmol. 2007;125(11):1531-1538. doi:10.1001/archopht.125.11.1531
2. Richter GM, Williams SL, Starren J, Flynn JT, Chiang MF. Telemedicine for Retinopathy of Prematurity Diagnosis: Evaluation and Challenges. Survey of ophthalmology. 2009;54(6):671-685. doi:10.1016/j.survophthal.2009.02.020.
3. Bartling, H., Wanger, P. and Martin, L. (2009), Automated quality evaluation of digital fundus photographs. Acta Ophthalmologica, 87: 643–647. doi:10.1111/j.1755-3768.2008.01321.x