**Study Title:** Evaluation of edge-based image quality assessment metric for AOSLO split detection images

**Purpose**

We have developed a new measure for assessing the image quality of adaptive optics scanning laser ophthalmoscopy (AOSLO) images by using automated edge detection to quantify the visibility of the cone mosaics in each image. The purpose of this experiment is to compare the consistency of our new quality measure against human quality evaluations of AOSLO split detection images. To evaluate this, we will use **X** montaged AOSLO datasets consisting of images that were withheld from the development process for the proposed quality assessment measure. From each dataset, we will identify Z location evenly distributed across eccentricities, with at least Y images that are directly overlapping at each location. We will then use the proposed quality assessment measure on these datasets to generate a ranking of the quality of the overlapping images from best to worst at each location. For evaluation, these rankings will be compared to manual rankings generated by a trained grader, and also rankings generated by two other previously published quality assessment measures: the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [1] and the Perception based Image Quality Evaluator (PIQE) [2].

**Hypothesis:**

We hypothesize that using the proposed automated quality assessment measure, we will be able to generate a ranking of image quality that is highly correlated to rankings generated by human graders, and on average performs better than rankings provided by BRISQUE or PIQE.

**Data**

**N** fully montaged AOLSO datasets with the split detector channel will be used to evaluate the proposed quality assessment measure. For each montaged dataset we will identify Z location with a minimum of Y overlapping images. These locations will be distributed as follows: 2 location at the fovea region, 6 sampled from arms, 3 locations sampled evenly from A degree to B degree along the temporal arm, and 3 locations sampled evenly from A degree to B degrees in the superior arm. Each location will serve as a different test point to evaluate the proposed automated quality assessment and ranking.

To provide a ground truth ranking at each location for comparison, two (one?) human graders will manually evaluate the set of images at each location and determine the best ranking of image quality within each set. This will be generated using a custom software that shows the graders each combinational pair of the overlapping images at a location, and then records the graders’ evaluation for which image in each pair has the better image quality. These pairwise comparison results will then be converted into an overall ranking for the images at each location using the Bradley and Terry Model [3].

**Method**

**Planned Experiment:**  For each image in each dataset, we will use the proposed quality assessment measure to produce an automated score. This measure will then be ordered to generate a within-set ranking for each set of overlapping images at each location. We will then compare the automated and manual rankings at each location to evaluate the efficacy of our automated quality assessment measure. The specific comparisons that will be made are described in the next section. For an external baseline comparison, we will also compute this evaluation for two other previously published automated image quality assessment metrics: the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [1] and the Perception based Image Quality Evaluator (PIQE) [2].

**Static Algorithm Parameters:** Our proposed quality assessment measure is generated using a standard Canny edge detection algorithm implemented in Matlab (v2020a). It has one parameter, which indicates the upper threshold for when an image feature is determined to be an edge (the lower threshold is always set to zero in our design). In this experiment we will test two approaches for setting this threshold parameter:

1. A static threshold of t=0.17 will be evaluated. This value was derived empirically from our development and pilot evaluation of the quality assessment measure. In this case, the output metric used as the quality assessment score is the fraction of the image that is detected as an edge using this static threshold.
2. An adaptive threshold will also be evaluated, where the algorithm optimizes for the minimum threshold required to fill 5% of the image with edges. In this case, the selected threshold itself is the output metric used as the quality assessment score from the algorithm.

**Dependent variables:** The variable being evaluated in this experiment is the ranking of the overlapping images at each location selected from the datasets, as determined by the quality assessment score from each algorithm.

**Analyses**

**Performance Measures:** We will assess each algorithm’s performance by evaluating the following measures:

1. **Pairwise specificity, sensitivity, and accuracy:** For each pair of overlapping images at each selected location, we have a manual evaluation of which image in the pair is of higher quality. We will use the automated quality assessment score to compare the same pairs of images and then evaluate how accurately the algorithm chose the correct image in each pair. We will calculate the overall specificity, sensitivity, and accuracy of this comparison for each automated quality assessment metric.
2. **Groupwise Spearman’s ranked correlation:** For each of the locations with overlapping images identified for testing, we will evaluate the Spearman’s ranked correlation coefficient between each automated and manual rankings of the set of overlapping images at that location.
3. **Qualitative assessment of full montages:** Using each automated quality assessment metric, we will reconstruct each AOSLO montaged dataset such that the image with the highest image quality measure score at a location will always be present at the top of the montage. For each dataset, a grader who is blinded to how each montage was created will then qualitatively assign an ordinal ranking between the top-level montages created using each algorithm and the original default montage.

All algorithm processing and evaluations will be run on the data by an individual who has no access to the ground truth manual rankings until the algorithm results are obtained.

**Comparisons:** We will test our hypothesis by comparing the performance of the proposed quality assessment measure relative to the inter-rater performance, and also against the performance of the two existing published image quality measures (BRISQUE and PIQE). We will test the proposed method using both the static and the adaptive threshold. We will look at:

1. The overall specificity, sensitivity, and accuracy of the pairwise comparison across all the datasets.
2. The mean and standard deviation of the Spearman’s rank correlation at each overlapping locations, across all datasets.
3. The mean qualitative ranking of each algorithm’s reconstructed top-level montage across all dataset (relative to each other and the original montage).