**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 cone mosaics in the 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** new AOSLO dataset consisting of images that were withheld from the development process for the quality assessment measure. We will then use the proposed quality assessment measure on this dataset to generate a ranking of the best quality to worse quality image for each overlapping region in each dataset. For evaluation, these rankings will be compared to manual rankings generated by a trained grader, and also rankings generated by X other previous published quality assessment measures.

**Hypothesis:**

We hypothesize that using the automated quality assessment measure, we will be able to generate a ranking of image quality that is highly correlated to rankings generated by a human grader.

**Data**

**N** fully montaged AOLSO datasets with the split detector channel will be used to evaluate the proposed quality assessment measure. Each montaged dataset will be divided into regions of overlapping images. Each discrete region will serve a different testing location to evaluate the automated quality assessment and ranking.

To serve as a gold standard, two (?) human graders will manually go through each dataset and determine the best ranking of image quality for the images in that data. This will be generated using a custom software that shows the grader each pair of overlapping images in the dataset, and the grader will select which image in the pair has the better quality. This pairwise comparison will then be used to determine an overall manual ranking of image quality at each of the discrete overlapping region for each grader.

**Method**

**Planned Experiment:**  For each image in each dataset, we will use the proposed quality assessment measure to produce an automated score. This measure is then used to generate a relative rank between each image in the dataset. We will then compare the similarities of the automated and manual rankings for each overlapping region in each dataset to evaluate the efficacy of our automated quality assessment measure. A gold standard will be created by comparing the ranking from the two different graders. For an external comparison, we will also compute this evaluation for N other previously published automated image quality assessment metrics: YYY, ZZZ, AAA.

**Static Algorithm Parameters:** Our 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 each overlapping region in each dataset, 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 on each of the datasets:

1. **Pairwise specificity, sensitivity, and accuracy:** For each pair of overlapping images in each dataset, we have a manual selection of which image in the pair is of higher quality. We will use the automated quality score to compare the same pairs of images and then evaluate how accurately the algorithm chose the correct image in each pair. We will report the specificity, sensitivity, and accuracy of this comparison for our proposed measure and existing established measure of image quality.
2. **Groupwise Spearman’s ranked correlation:** For each overlapping section of **N** or more images in each dataset, we will evaluate the Spearman’s ranked correlation coefficient between each automated and manual rankings of the images.
3. **Qualitative assessment of full montages:** Using the automate rankings we will reconstruct the full montage for each dataset. A trained grader will then qualitatively assess the overall quality of the top image level of each montage for the purpose of future analysis. (*Needs more thought.*)

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 gold standard inter-rater performance, and also against the performance of the existing published image quality measures. 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 across all overlapping regions.