Measuring Attention in Online Meetings using Computer Vision

Tristram Dacayan and Daehan Kwak
School of Computer Science and Technology, Kean University, Union, NJ, 07083

Keywords: Computer Vision, Gaze Estimation, Object Detection, Object Tracking, Attention

Whether it is a meeting, presentation, or lecture, an audience's attention is one of the fundamental factors determining whether the objective is understood. Assuming this audience is not as knowledgeable as the presenter, participants would need to pay attention in order to understand the point a speaker may be attempting to put forth. Although it is not a new concept, attempting to assess an audience's attention while presenting or teaching a lecture could be difficult and time-consuming for the presenter and awkward for the attendees. However, if it were possible to do so without downsides, knowledge about when and who becomes uninterested or unengaged could help presenters and the audience identify the faults in either the presentation or themselves.

In this research, the goal is to utilize computer vision libraries and pre-trained object detection models to create a tool that could calculate the average rate of attention during web conferences. While data from real-life meetings could be a more accurate source of data, the persistence of COVID-19 causes real-life meetings to become scarce and give way to the rise of online meetings. This goal is accomplished by developing an application using Python since many viable computer vision applications are created in that language.

To accurately calculate the collective attention rate, the program employs the usage of several different Computer Vision libraries such as OpenCV, a computer vision library consisting of all the primary tools needed for the rendering and tweaking of the imputed video, and YoloV3 and TensorFlow, both of which are required for their object detection capabilities. All three libraries are used to detect all the people within a single frame of the video. Once identified, DeepSort, a deep learning algorithm generally used for object tracking, calls a method to keep track of the people in the frame by assigning them a unique ID number and storing them in a separate container. Because the goal is to calculate the attention rates of every person in the video, DeepSort provides the much-needed organization required for the latter half. Once detected, a separate image is created based on the coordinates generated by the object detection libraries and gets stored within their respective folder. However, if that image is the first instance of the detected person, a new folder is created, and the image is stored there. Once the video is finished processing and detecting all the people, the program then iterates through each collection of photos taken and compiles them into a video that can be viewed individually within the output folder.

Because each video has at least one detected person in the frame at all times, the video is sent through a pre-trained gaze-estimation module called PTGaze. Essentially, this module estimates the trajectory of a person's gaze and head orientation based on the ETH-XGaze dataset through the use of several deep learning algorithms. Sending the videos through the module creates an overlay to visualize the head orientation and estimated gaze, thus allowing the ability to record the duration of a person's head and gaze within a specific region. For instances of inattention, a timer is started when a person's gaze or head orientation exceeds a certain threshold, granting them a few seconds before determining that they are not paying attention. In the instances where the person leaves and comes back into the frame, the time they disappear is recorded and included in the final calculation. Once the video is processed, the resulting attention rate is computed by dividing the total time recorded, paying attention by the original length of the recording. Each attention rate is then combined to determine the average rate of attention of the meeting.

After all the calculations and the video processing, a user could view each output video with the attention rate and timestamps of significant disinterest to the side. From the output, the user could potentially determine when most of the attendees lost interest in the meeting and stopped paying attention. This knowledge allows the user to reflect on crucial points of disinterest and improve their presentation for the future. Inversely, this data could also be used to determine a correlation between a group or individual's attention rate and their overall efficiency or effectiveness, which could help identify problems that may have been hard to notice before its usage. In addition, this system has much more potential in face-to-face classrooms, where knowing a student is paying attention can be useful for teachers since there is more concrete correlation between their attention span and their performance.