Towards integrating eye gaze tracking into a multimodal dialog agent for remote patient assessment

Daniel Tisdale, Jackson Liscombe, David Pautler and Vikram Ramanarayanan

Abstract We demonstrate a prototype that integrates automated eye gaze tracking into an already existing multimodal conversational platform for remote patient assessment and monitoring (NEMSI; short for "NEurological and Mental health Screening Instrument"). The platform engages patients in an interactive dialog session and guides patients through several spoken, orofacial, cognitive, and gaze tasks inspired by clinical protocols. Novel additions to the dialog protocol include a selection of exercises that have been widely used in oculomotor pathology research as well as clinical practice, including: smooth pursuit, saccade, free image exploration, directed image exploration, and the congruent and incongruent Stroop tests. Furthermore, the prototype automatically computes eye-gaze metrics in addition to speech, facial, linguistic, and motoric metrics relevant to the assessment of their overall neurological and mental health. Finally, we report on internal testing to validate the accuracy of real-time eye gaze software and metrics shown to be of use in clinical research.

1 Introduction

Eye tracking is the process of measuring either the point of gaze (where one is looking) or the motion of an eye relative to the head (1). Scientific exploration of eye gaze has been investigated since as far back as the 19th century (2). Today, there are several eye gaze tracking software products that gather data in completely non-invasive ways without any equipment needed other than a computer or smart phone[1]. This relatively recent technology has the potential for remote monitoring of patient

Daniel Tisdale and Jackson Liscombe equally contributed to this paper.
Modality.AI, Inc.
e-mail: vikram.ramanarayanan@modality.ai

[1] https://www.bitbrain.com/blog/eye-tracking-devices

biomarkers to evaluate the effectiveness of treatment of neurological and cognitive impairments. Indeed, abnormalities in eye gaze metrics have been clinically validated for many diseases, such as multiple sclerosis, AIDS dementia complex, antisocial personality disorder, autism spectrum disorder, schizophrenia, psychosis, dyslexia, eating disorders, social anxiety disorders, attention deficit hyperactivity disorder, fetal alcohol spectrum disorder, Parkinson's disease, and bipolar disorder (3; 4; 5). Therefore, there is a well-established relationship between eye gaze data and cognitive and neurological functioning. Several tasks (and associated metrics) derived from in-clinic eye gaze assessment protocols can capture this relationship and its breakdown. For instance, a saccade is the rapid movement the eyes do simultaneously to change the line of sight. Smooth pursuit eye movements are the voluntarily tracking performed when stabilising gaze on a moving visual target. Fixations are the stationary states of the eyes during which eye gaze is held upon a specific location in the visual scene. Fixations can be furthermore incorporated into saliency metrics based on models of human attention to certain locations in a video or picture. Finally, the entire path of a gaze sequence, or scan path, for a particular task can be considered as either a shape in and of itself or as input into machine learning algorithms. Such metrics derived from eye gaze movements have been shown to correlate with several cognitive and neurological disorders, both degenerative and developmental. In the following sections, we demonstrate how we have incorporated non-invasive eye gaze tracking software into a cloud-based multimodal dialog system for remote patient assessment and monitoring.

2 Interaction Flow

2.1 NEMSI

NEMSI (NEurological and Mental health Screening Instrument) is a multimodal conversational platform for remote patient diagnosis and monitoring, which extracts a variety of biomarkers after engaging patients in an interactive dialog session (6). The obtained biomarkers have been shown to be useful for a number of neurological conditions and can be visualized in a user-friendly dashboard for further analysis (7; 8; 9; 10; 11; 12; 13). The conversational assessment protocol can include a customizable subset of following tasks, depending on the nature of the disease in question: an oral motor exam, sustained phonation exercises, diadochokinesis exercises (rapidly repeated syllables), read speech, including isolated words, sentences and read passages, spontaneous speech prompts, spirometric exercises such as exhalation and coughing, picture description, emotional state elicitation, and other cognitive tasks such as recalling previous words or numbers.

2.2 Eye Gaze Tracking

We use Webgazer.js[2] (14) as our eye gaze tracker in NEMSI, which is licensed under GPLv3. WebGazer.js has two key components: a pupil detector that can be combined with any eye detection library and a gaze estimator using regression analysis informed by user interactions. The eye detection library we use is MediaPipe Face Mesh[3] (15). Webgazer.js was originally designed for use in evaluating user interaction with websites and, as such, by default uses feedback from user mouse movements for continuous gaze calibration. Because many of our users suffer from diseases that may affect motor control and impair their use of a mouse, we have turned off this Webgazer.js feature and instead start with just one mouse-related calibration task, described below.

The NEMSI virtual agent can engage the user in the following tasks:

- A calibration task in which the user is asked to click 9 circles five times each while looking directly at the circle they are clicking. The circles are arranged at the border of their browser screen and the results inform the eye gaze tracker to better estimate user gaze position for the remainder of the session (Figure 1.)
- Extreme vertical and horizontal eye gaze tasks in which the user is asked to gaze as far to the left, right, up, and down as they can, both slowly and rapidly.
- Smooth pursuit tasks in which the user is asked to follow a moving circle with their eyes while it moves in either a line or a circle.
- A saccade task in which users are asks to direct their gaze to dots that appear briefly on the screen in random locations.
- Modified congruent Stroop Color and Word tests (SCWT) (16) in which the user is presented with a matrix of color words (e.g. “green”,“purple”, etc.) whose ink color matches their semantic meanings and is asked to either read the words in order or to find a specific word, as directed by the automated agent (Figure 2).
- Modified incongruent SCWTs identical to the ones above, but in which the ink color of the words do not match their semantic meanings (e.g. the word “brown” with an ink color of pink). See Figure 3.
- Picture exploration and description tasks in which users are shown pictures and asked to explore them visually and to describe what they see, or are instructed to find specific items within the pictures. See Figure 4.

3 Analytics and Verification

This section presents proof-of-concept results for showing the reliability of the data collected from the eyetracking tests. The first thing to verify was that our observed error rate matched that reported by WebGazer.js developers. We designed a “shrinking dot task” in which internal testers were asked to use their eye gaze to follow

[2] `https://webgazer.cs.brown.edu/`

[3] `https://google.github.io/mediapipe/solutions/face_mesh`

Fig. 1: Calibration

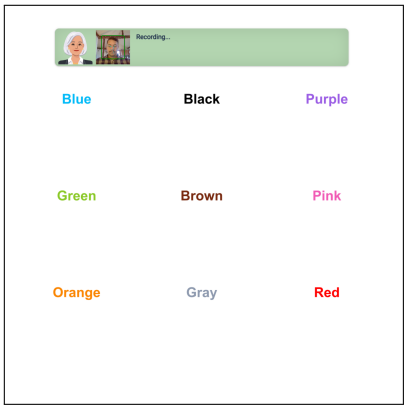

Fig. 2: Congruent SCWT

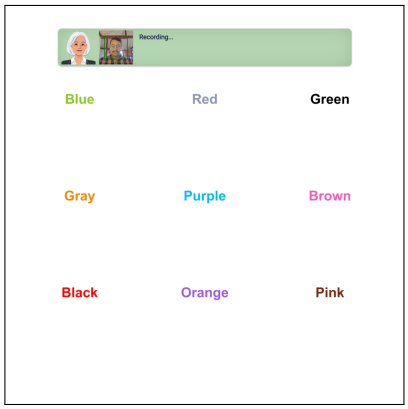

Fig. 3: Incongruent SCWT

Fig. 4: Picture Exploration

a smoothly moving dot as it moved and bounced off the edges of their computer screens. Over time, the size of the dot grew progressively smaller. The movement of the dot around the screen allowed us to evaluate the model's accuracy at different coordinates on screen, while the shrinking size helped us compute the accuracy and precision of the eye tracking software. We found the mean accuracy for the task fell below 50% once the dot was smaller than 180 pixels (px), on average. This result falls within the mean error reported in (14) of 210.6px (SD=86.3px). This error of 180px has informed our task user interface design in that set fixation targets at least 200px apart from one another.

We also analyzed eye tracking data of 13 healthy controls who completed all aforementioned tasks. Users had screen sizes of varying dimensions, as expected. For this reason, all eye gaze coordinates were normalized to the range [0,1] based on their respective screen dimensions. Average screen size of users was 858x1610px. Figures 5 and 6 show two examples of the x coordinate eye gaze data obtained from a single user in the smooth pursuit and saccade tasks, respectively. From vi-

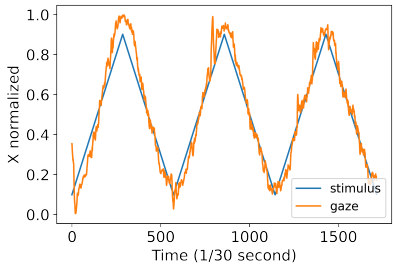

Fig. 5: Smooth Pursuit Task Example (X coordinates)

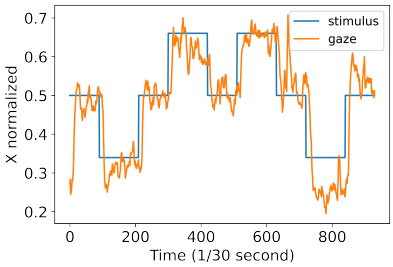

Fig. 6: Saccade Task Example (X coordinates)

Table 1: Eye Gaze Prediction Errors

Task	Error X (%)		Error Y (%)	
	Mean	SD	Mean	SD
Smooth Pursuit Line	8.8	3.7	23.4	6.6
Smooth Pursuit Circle	11.2	4.9	21.2	8.5
Saccade	7.5	5.3	19.6	4.1
Congruent SWCT	13.7	10.8	23.6	10.6
Incongruent SWCT	14.2	12.6	21.1	8.3
Image Saccade	14.8	12.8	20.7	11.8

sual observation, the accuracy is promising. Note that many metrics can be successfully computed from less-than-perfect fixation accuracy because many are statistical functionals, derived from overall gaze span shape or movement velocities. For example, in Figure 6, though the fixations points are sometimes off, clear saccades can be seen from one fixation point to the next, including the time taken to respond to a new target. In addition to visual inspection, we also computed fixation error over all tasks and subjects. Table 1 lists average x and y displacement error per task, as percent of screen size. Given average screen size, and considering 180 pixel average error, expected displacement error for healthy controls should be $x \leq 11.2\%$ screen width and $y \leq 21.0\%$ screen height. Most are within this expected range and for those that are not, future investigation is warranted. All in all, these data and analyses on healthy controls provides us with a successful proof-of-concept towards our next step: investigating the feasibility and utility of deploying this technology to analyze data from patients with cognitive and neurological disorders.

4 Conclusion

We have demonstrated how to incorporate the eye gaze modality into NEMSI, a multimodal conversational platform for remote patient diagnosis and monitoring, which extracts speech, facial, cognitive, respiratory and now eye-gaze-based biomarkers while engaging patients in an interactive dialog session. Further information about NEMSI can found at `www.modality.ai`.

References

[1] C. H. Morimoto and M. R. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer vision and image understanding*, vol. 98, no. 1, pp. 4–24, 2005.

[2] M. Płużyczka, "The first hundred years: a history of eye tracking as a research method," *Applied Linguistics Papers*, vol. 4/2018, pp. 101–116, 12 2018.

[3] M. Vidal, J. Turner, A. Bulling, and H. Gellersen, "Wearable eye tracking for mental health monitoring," *Computer Communications*, vol. 35, no. 11, pp. 1306–1311, Jun. 2012.

[4] A. Wolf and K. Ueda, "Contribution of Eye-Tracking to study cognitive impairments among clinical populations," *Front Psychol*, vol. 12, p. 590986, Jun. 2021.

[5] P.-H. Tseng, I. G. M. Cameron, G. Pari, J. N. Reynolds, D. P. Munoz, and L. Itti, "High-throughput classification of clinical populations from natural viewing eye movements," *J Neurol*, vol. 260, no. 1, pp. 275–284, Aug. 2012.

[6] D. Suendermann-Oeft, A. Robinson, A. Cornish, D. Habberstad, D. Pautler, D. Schnelle-Walka, F. Haller, J. Liscombe, M. Neumann, M. Merrill *et al.*, "NEMSI: A multimodal dialog system for screening of neurological or mental conditions," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 245–247.

[7] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, and J. R. Green, "Speech as a biomarker: Opportunities, interpretability, and challenges," *Perspectives of the ASHA Special Interest Groups*, pp. 1–8, 2022.

[8] M. Neumann, O. Roesler, J. Liscombe, H. Kothare, D. Suendermann-Oeft, J. D. Berry, E. Fraenkel, R. Norel, A. Anvar, I. Navar, A. V. Sherman, J. R. Green, and V. Ramanarayanan, "Multimodal dialog based speech and facial biomarkers capture differential disease progression rates for ALS remote patient monitoring," in *Proceedings of the 32nd International Symposium on Amyotrophic Lateral Sclerosis and Motor Neuron Disease 2021*, Online, December 2021.

[9] H. Kothare, M. Neumann, J. Liscombe, O. Roesler, W. Burke, A. Exner, S. Snyder, A. Cornish, D. Habberstad, D. Pautler, D. Suendermann-Oeft, J. Huber, and V. Ramanarayanan, "Statistical and clinical utility of multimodal dialogue-based speech and facial metrics for Parkinson's disease assessment," in *Proceedings of Interspeech 2022*, Incheon, Korea, September 2022, pp. 3658–3662.

[10] H. Kothare, V. Ramanarayanan, O. Roesler, M. Neumann, J. Liscombe, W. Burke, A. Cornish, D. Habberstad, B. Kopald, A. Bai, Y. Markiv, L. Cole, S. Markuson, Y. Bensidi-Slimane, A. Sakallah, K. Brogan, L. Lampinen, S. Skiba, D. Suendermann-Oeft, D. Pautler, and C. Demopoulos, "Atypical speech acoustics and jaw kinematics during affect production in children with Autism Spectrum Disorder assessed by an interactive multimodal conversational platform," in *Proceedings of the 8th International Conference on Speech Motor Control (SMC) 2022*, Groningen, the Netherlands, August 2022.

[11] M. Neumann, O. Roesler, D. Suendermann-Oeft, and V. Ramanarayanan, "On the utility of audiovisual dialog technologies and signal analytics for real-time remote monitoring of depression biomarkers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) Workshop on NLP for Medical Conversations 2020*, Online, July 2020.

[12] H. Kothare, M. Neumann, J. Liscombe, O. Roesler, D. Habberstad, W. Burke, A. Cornish, L. Arbatti, A. Hosamath, D. Fox, D. Pautler, D. Suendermann-Oeft, I. Shoulson, and V. Ramanarayanan, "Assessment of atypical speech in Multiple Sclerosis via a multimodal dialogue platform: An exploratory study," in *Proceedings of the 8th International Conference on Speech Motor Control (SMC) 2022*, Groningen, the Netherlands, August 2022.

[13] A. Khan, S. Prokop, S. Bashir, J.-P. Lindenmayer, B. Insel, D. Pautler, D. Suendermann-Oeft, C. Yavorsky, and V. Ramanarayanan, "Reliability, validity and internal consistency of multimodal AI based facial and acoustic biomarkers of negative symptoms in schizophrenia," in *Annual Meeting of the Schizophrenia International Research Society (SIRS) 2022*, Florence, Italy, April 2022.

[14] A. Papoutsaki, P. Sangkloy, J. Laskey, N. Daskalova, J. Huang, and J. Hays, "WebGazer: Scalable webcam eye tracking using user interactions," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*. AAAI, 2016, pp. 3839–3845.

[15] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," *CoRR*, vol. abs/1907.06724, 2019. [Online]. Available: http://arxiv.org/abs/1907.06724

[16] F. Scarpina and S. Tagini, "The stroop color and word test," *Frontiers in Psychology*, vol. 8, 2017. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00557