FocalVid : Facilitating Remote Studies of Video Saliency

Sahand Shaghaghi*, Bryan Tripp*, Chrystopher Nehaniv*† Alexander Mois Aroyo†, and Kerstin Dautenhahn†*
*Department of Systems Design Engineering
†Department of Electrical Engineering
University of Waterloo, Waterloo, Canada
Email: {s2shagha, bptripp, cnehaniv, aaroyo, kdautenh}@uwaterloo.ca

***Abstract*—Humans selectively use only a small fraction of the vast sensory data that arrives at their receptors. In vision, eye movements are an important part of this selection process. Eye movements arise from interacting bottom-up and top-down factors, including social factors, and they provide important information about cognitive processes. Tracking eye movements often requires very specialized hardware that is suitable for laboratory based studies, but less practical for online studies with remote participants. Recently, a proxy for eye movements was introduced, which facilitates large online studies of overt attention. In this approach, participants' mouse-clicks reveal parts of a static image sequentially. Mouse-click locations can be recorded accurately and reliably, without the need for a calibration procedure, and mouse-click locations were found to correlate strongly with gaze locations. However, while eye movements are often studied using static scenes, they are also affected by motion cues, and by more complex task-related dynamics. To facilitate the online study of such influences, we adapted the mouse-based approach to dynamic scenes, by continuously recording the location of the mouse cursor, and continuously revealing only part of the display surrounding the cursor. While our platform has been developed primarily to support large, remote video saliency studies, the same approach could be used to study overt attention, e.g., in computer games, or to study more complex interactions, such as co-operative tasks performed over video chat. This paper describes our platform, FocalVid, which will be made open-source on acceptance.**

Keywords–eye tracking; visual saliency; video saliency; mouse-contingent interface; Web design.

I. INTRODUCTION

In human-human interactions, gaze behavior and visual preference choices are crucial since they influence and regulate the dynamics of the interaction. Similar dynamics are also present in Human-Computer-Interaction (HCI) scenarios. It is important to fully understand the dynamics of gaze interactions in these scenarios. Increasingly online studies are conducted with remote participants, in addition to laboratory in-person studies. Thus, it is timely to explore methodologies which facilitate seamless recording of gaze and visual saliency for a broader participant base that can be used across a variety of hardware and operating systems. The presented methodology makes it possible to explore remote human-human or human-robot interactions by allowing the researchers to observe and record participants' visual selection data for various scenarios involving video saliency, overt attention scenarios and co-operative remote tasks. Such gathered data is valuable not only in the field of HCI, but also in fields of computer vision, ergonomics, user experience design and Human-Robot Interaction (HRI) to name a few since it will ultimately enable researchers to make design choices based upon gaze and human visual preferences.

Eye movements and gaze patterns influence the dynamics of social interactions. Eye movements/fixations and gaze patterns are intertwined. Eye movements lead to instances of gaze, but not all eye movements necessarily lead to socially meaningful gaze instances. Theories [1] and models that have been discussed in the literature [2] attempt to make sense of these gaze patterns in social interaction, exploring the intricacies of gaze behavior in mutual gaze, joint attention, dyadic and triadic interactions. There is a need for new tools to further explore different aspects of these models and theories. Here, we introduce FocalVid as an effort towards this end. FocalVid facilitates recording of visual attention patterns of articipants while viewing videos in remote settings.

The correlation between visual selection and hand movements [3], and also the close correlation between gaze and cursor locations [4]–[8] have been established previously. This chain of correlations is the main rationale behind (computer) mouse-contingent methodologies, including FocalVid, which utilizes participants' controlled cursor movements on a visual canvas to record proxies for gaze behavior. Up to now, eye movement tracking has been conducted predominantly using expensive equipment in controlled laboratory environments to record participants' eye movements [9], which is limited in reach and typically cannot be used for remote, e.g., crowd-sourcing studies [10]. FocalVid departs from this approach by making use of a mouse-contingent methodology, which allows the participants' presence in a closed interaction loop involving the participant and the scenario unfolding in a video that they observe, facilitating broader participant reach due in remote participation.

The presented platform is not only useful in the study of gaze, but it is also useful in a multitude of experimental scenarios such as: video saliency studies, video game usability studies, and platform usability studies. The rationale that FocalVid is based on closely correlates with the concept of visual saliency [11], exploring the utilization of the bottom-up saliency concept [12]–[14] to establish a relationship between visual features and gaze directions. The field of visual saliency is interested in points of attention in 2D and 3D scenes [15]. Such a relationship then makes it possible to evaluate the findings gained with FocalVid against available findings from the field of human visual saliency studies which could be used to evaluate and benchmark the presented system. Our approach extended the methodologies designed by Kim et al. [4], Jiang et al. [16], and Jansen et al. [17] through the redesign and extension of those approaches with the addition of video playback. Such context has not been explored in detail previously, and is a novel contribution of the present work. As such, the main contribution of this study is the presentation of a system that would make it possible to record participant visual selection for any given video scenario.

The remainder of this article is structured as follows. Section II discusses related work, followed by a description of the design and implementation of FocalVid (Section III). Results of initial system tests are presented in Section IV and discussed in Section V. Section VI concludes the article.

II. Related Work

Related works are categorized into thematic areas: gaze, visual perception and visual saliency which has close ties with visual perception. Efforts relating to mouse-contingent methodologies are also reviewed which directly relates to the system designed in this study.

A. *Gaze, Visual Perception and Eye-tracking methodologies*

There is a rich history of research in the field of vision which deals with eye movements, gaze, and visual perception. There have been various attempts at the recording of eye movement in the past 60 years [18]. These attempts initially used more intrusive apparatus and since have moved toward less intrusive solutions [19, p. 9]. Initially, eye movements recording devices needed to be connected to the sclera such as the apparatus designed by Yarbus [20]. He developed an apparatus to accurately record eye movements using suction caps which attach to the sclera. Eye movement recording solutions have generally become less intrusive [9], often using cameras and infrared illumination to improve contrast between the pupil and surrounding tissue, but they still typically require expensive specialized hardware. There have been two new approaches that have challenged this tendency. Both of these approaches are moving in the direction of more broadly accessible methodologies and systems:

1) Appearance-based tracking: This approach attempts using visible-light cameras to track participant's eye movements. This method of eye tracking makes it possible to conduct remote studies using embedded Webcams in participants' personal computers [21]–[23].
2) Mouse-contingent tracking: This approach is the main focus of this study. These methodologies utilize cursor location to determine participant's visual selection. We elaborate on those issues in more detail in Section II-C.

B. *Visual Saliency*

Visual saliency refers to ”bottom-up factors that highlight image regions that are different from their surroundings” [24, p.1], which make these regions in the visual field of interest to viewer. As such, there is a connection between visual saliency and eye movements: If a feature is highly salient, then there is a high likelihood of eye movement patterns whose trajectory dwells in that feature's spatial distribution within the field of view. One of the goals of the research field of visual saliency is the creation of models that could anticipate these highly salient features. These models are either hand-crafted [12][14][25][26] or, recently, deep-learning based [27]–[29]. Visual saliency explores detection for both static and dynamic scenarios [15]. Static scenarios mainly deal with static images and dynamic scenarios mainly deal with videos. Even though there are commonalities between the two settings, observation patterns differ for these two instances: The duration of viewing is different in the two, and dynamic cases have the extra element of motion with respect to the observer.

A necessity in the field of visual saliency is benchmark annotated data which models could be trained and tested against. This could either be in the form of eye-tracking benchmark data or crowdsourced benchmark data. Initially, it was customary to use smaller eye-tracking datasets for the training of hand-crafted models, but with the move to deep learning based models, larger databases are needed. Methodologies such as SALICON [16] become of use in these cases since such a methodology allows for the gathering of benchmark data from a broader participant base using services like Amazon Mechanical Turk [22]. As such, the creation of systems that make this type of data gathering possible would be of value, especially relating to video saliency where such solutions still are not readily available.

C. *Mouse-contingent Methodologies*

A current focus of the field of eye-tracking is the development of methodologies for more efficient collection of eye movements [4, p.4]. Such methodologies, including mouse-contingent ones, would then enable researchers to conduct experiments with a broader reach. Mouse-contingent methodologies make use of computer mouse and cursor location, which can be reliably recorded, and encourage co-origination of eye and mouse movements in various ways. These methodologies enable researchers to conduct saliency and usability studies through online platforms, which can be a significant advantage over time-consuming in-person laboratory experiments with expensive specialized software and hardware. Mouse-contingent methodologies have their roots in psychology studies. A seminal study investigates the use of bubble-shaped visual windows [30] for the evaluation of recognition tasks. Both visual windows and visual scotomas [31] have been explored extensively in the field of visual perception psychology which has lead to the moving-window approach in the field of HCI. An early example of such an approach is RFV viewer [17]. There have been some renditions and improvements of this model ever since [32][33] with the newest additions being SALICON [16], and Bubbleview [4]. These platforms are all designed with static images in mind and hence this creates the need for the design of a platform that could be utilized for video saliency.

III. Method

In this section, the FocalVid system design (Figure 1) is detailed. First, the fundamental moving visual window components of the system are detailed (Section III-A), followed by details regarding system interface (Section III-B) culminated by system implementation details (Section III-C).

A. *Controlling Video Visibility via Mouse Movements*

Here, much like BubbleView [4], the concept of bubbles was investigated as a base for the proposed platform [30][34]. This concept was then expanded upon by the incorporation of opacity, visual moving windows, and addition of the ”graded regions of stimuli” [17] in the form of concentric circles. Our approach makes use of opacity instead of blurring used by Kim et al. [4].

Our platform displays a video behind a semi-transparent layer. This layer is nearly opaque over most of the video frame, but it is more transparent around the mouse-cursor

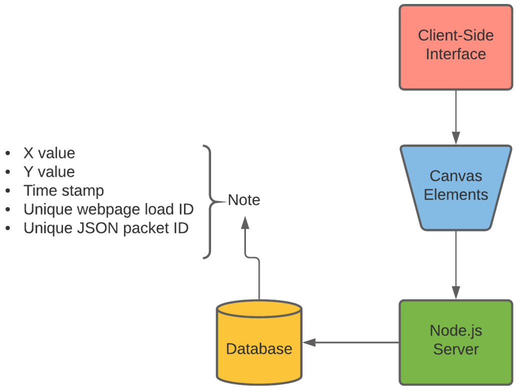


Figure 1. FocalVid system diagram.

location. The transparency is greatest within a small circle around the cursor location, and increasingly opaque in circles of increasing radius (Figure 2). This pattern reflects humans' higher visual acuity closer to the fovea, and approximately radially symmetric [35] decreases in acuity approaching the periphery. The result is that the video can be seen most clearly when gaze is centred on the cursor position. Participants must co-ordinate mouse and gaze to clearly see any part of the video. Importantly, partial transparency farther from the cursor allows detection of salient cues, but greater transparency at the centre encourages correspondence between gaze position and cursor position.

In our initial implementation described in this article, the circle sizes and transparencies are chosen by trial and error, with the goal of maximizing correspondence between natural gaze and mouse-cursor position. In the example of Figure 2, we have intentionally made the circles too large, to illustrate that this makes it too easy to see much of the scene without having an intention to move the cursor. To reduce the visual complexity of the display, as well as the complexity of the parameter space, the radii of circles always increase in uniform steps. We use a small number of circles, because displaying greater numbers of circles is more computationally demanding. Note, circle sizes and transparencies can be adjusted for the implementation of specific experiments, and might depend, among other factors, on the content displayed in the videos and/or associated research questions with regard to the type of data researchers intend to collect.

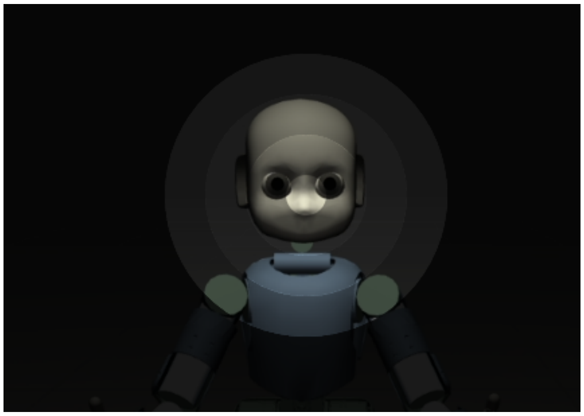
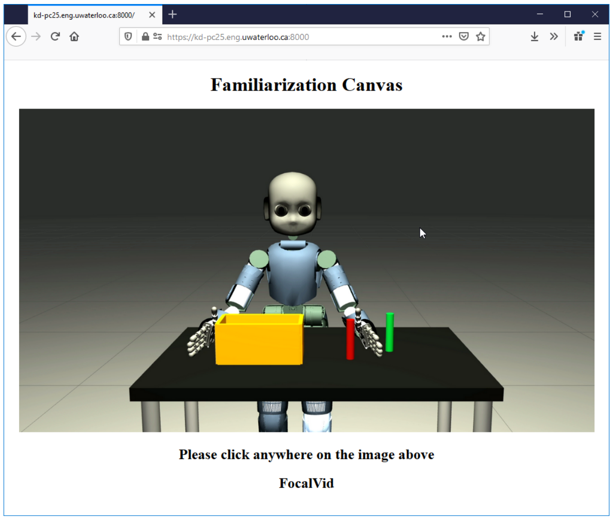


Figure 2. Details relating to radial bubbles.


kd-pc25.eng.uwaterloo.ca:8000/
https://kd-pc25.eng.uwaterloo.ca:8000

Familiarization Canvas

Please click anywhere on the image above

FocalVid

Figure 3. Get ready page: Here participants are asked to move their cursors inside the boundaries of the present image and then click in order for the experiment to begin. This is done so that participants' cursor is located on the canvas when the experiment begins.

B. *Experiment Interface*

The interface is Web-based, and implemented with HTML and JavaScript. The first page shows a short video (unrelated to the main experiment) and allows participants to freely practice using the interface without being evaluated.

The second page displays the first frame of the experimental video, in order to familiarize the participants with the general scene and points of interest, until the participant is ready to begin the experiment (Figure 3). The participant must move the mouse into the video canvas and click to begin. This ensures that the participant's attention is on the canvas when the experiment begins.

The third page presents the main experiment interface (Figure 4). This page includes one HTML canvas, with two layers. The bottom layer displays the video, and the top layer contains the semi-transparent overlay (Figure 2).

When the main experiment page has finished loading, a timer starts, and the video plays automatically. Cursor positions are detected via JavaScript "mousemove" events, time-stamped using the timer, and stored in a database. "mousemove" events report updates to the cursor location when the cursor is moved. When the video ends, the interface loads a final "thank you" page, and the experiment is complete.

C. *Implementation Details*

The implementation uses JavaScript, both in the client browser, and in a Node.js environment on the server. The client-side interface makes a secure HTTPS connection with a Node.js server, using the fetch method. The system continually transmits timestamped cursor positions to the server. Each instance of the experiment is assigned a unique ID, which is bundled with the cursor data, allowing multiple experiments to run in parallel.

The system stores data using NeDB [36], a widely-used NoSQL database that is compatible with Node.js. NeDB stores data in a simple JSON text file. The database stores x and

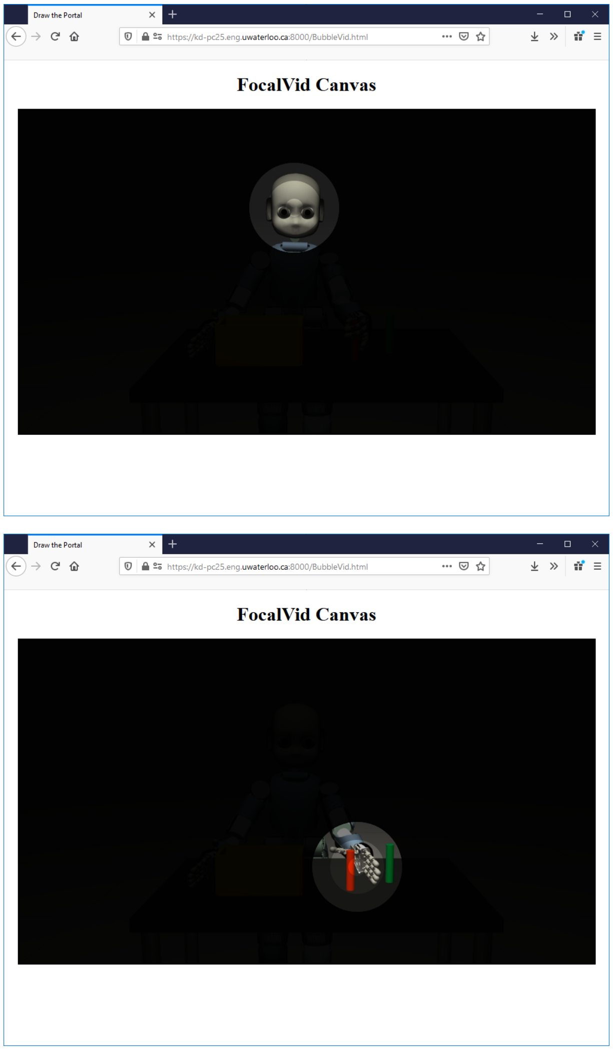

Draw the Portal
https://kd-pc25.eng.uwaterloo.ca:8000/BubbleVid.html
FocalVid Canvas
Draw the Portal
https://kd-pc25.eng.uwaterloo.ca:8000/BubbleVid.html
FocalVid Canvas

Figure 4. The main experiment page. Here participants' cursor is centered on the most inner circle of the co-centered circles. With the movement of the cursor, positioning of the concentric circles is altered, bringing into focus different elements in the video. In the upper figure the robot's head is focused on. In the lower figure the robot's left hand is focused on.

y coordinates of the cursor, timestamps of the coordinates, a unique identifier for each Webpage load, and a unique identifier for each JSON object. These JSON objects are continuously logged into the database. The unique identifier is created using the crypto API in the client-side platform. This API accommodates some cryptographic methods including random key generation.

IV. Results

To technically test the system and confirm that it works as expected in a realistic experimental scenario, with a variety of hardware and browser software, the authors served as 'participants' in a mock experiment. We used FocalVid to view a video (with the duration of 20 seconds) of a simulated iCub robot [37] performing a sorting task. We viewed the same video twice, once focusing on the robot's face (task 1), and the other time focusing on the robot's left hand, which was moving objects into a box (task 2). iCub is an open-source humanoid robot broadly used by researchers.

Figure 5 shows an example trajectory (horizontal and vertical mouse positions versus time stamp) from the task 2. The cursor pauses from time 3064 (ms) to 5232 (ms), at a location on the robot's left hand.

Recorded mouse-contingent data was then used to produce a heatmap (Figure 6). Heatmaps are a visualization method used to visualize density distributions of points in 2D and 3D settings [38]. This then further confirms that the system is recording the proper mouse-contingent data at the proper timestamp. Here, a random frame from the viewed video was chosen for analysis. A heatmap of the recorded cursor data belonging to all viewing instances was then produced using the Seaborn library [39] in Python. This heatmap presentation illustrates the kernel density estimation for the logged cursor points in relation to the video frame. As seen in Figure 6, there are two major areas of interest aligning with the robot's face and the robot's left hand which is moving objects into the box. Outlier data points can be observed as well.

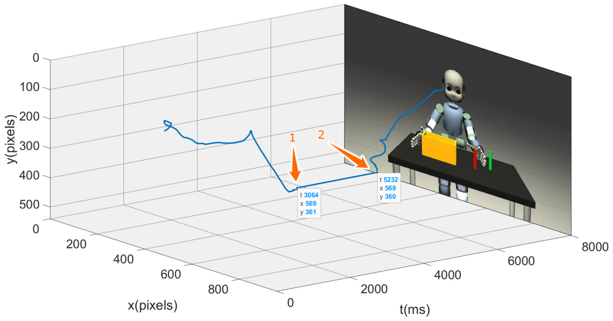


Figure 5. Data log visualization of recorded mouse-contingent cursor locations using FocalVid. The axes of the 3D graph represent x, y pixel coordinates in the video canvas and t, the time in milliseconds since the start of the video. Note that x & y refer to mouse locations in relation to the video frame and t refers to the timestamp belonging to that cursor location recording. An example of dwell of the cursor and hence the clearest visual region on the robot's left hand could be seen in this instance.

As another element of the technical evaluation of the system, to test the participant cursor and interface correspondence with the logged data, an additional set of recordings was made

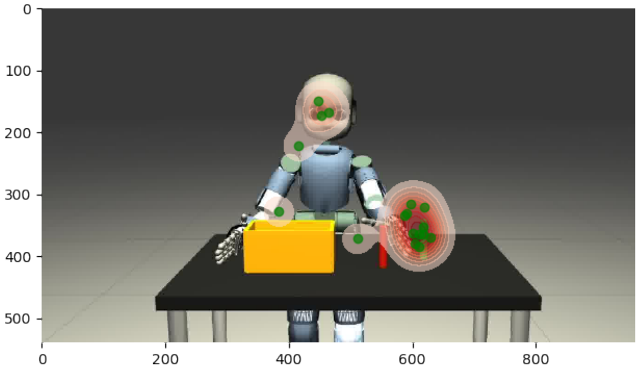


Figure 6. Heatmap of the mouse-contingent points recorded through the first two experiments using the authors. Here the data adjacent to a single frame was processed. Two major clusters could be seen, one belonging to the face area of interest (AOI) and the other belonging to the left hand AOI. Notice that the hand AOI does not fully align with the hand location, possibly due to presence of other highly salient features in the vicinity of hand, etc.

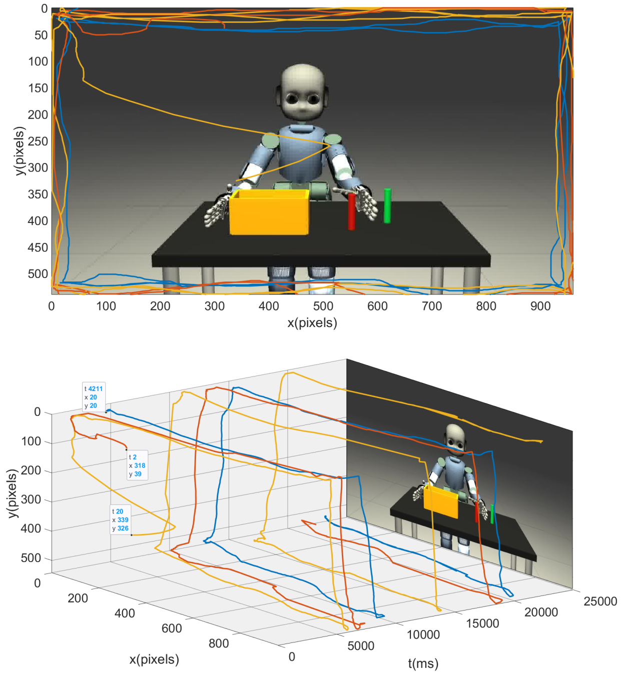


Figure 7. Three samples of the ten ”4corners” experiment iterations conducted, plotted side by side. Note that the mouse movements could start immediately at the beginning of the experiment or with a delay, depending on participants' initial reaction time.

using the authors while viewing the same iCub video, assigned with a new task (‘4corners’): The authors were instructed to move their mouse in a clockwise direction, attempting to get as close as possible to the four corners of the video frame in sequence starting from the upper left hand corner followed by the upper right hand corner and so forth. Example trajectories can be seen in Figure 7. The corner point cursor data for the ten recorded experimental instances were then extracted from this plot. Note that there was some variance around the targets, as is expected generally. In this case the variance was self-selected, as the authors were not given instructions about how to trade off speed and accuracy, or any other task constraints.

The final experiment conducted was the single point experiment. The rationale for this experiment was to assess the consistency of the system's performance across different operating systems, different browsers and screens. What was tested was if dwell on a specific point in the video frame would be accurately transmitted and recorded in the database. Authors were tasked with an experimental setup in which they were instructed to hover the mouse over the upper left hand edge of the yellow box present in the iCub video scenario. In order to achieve precision, cursor pointer was enabled in the interface. Results were positive attesting to the correct transmission of the data to database, with single cursor location being recorded for the entirety of the time cursor was positioned on the yellow box's edge.

V. Discussion

This presented platform makes use of *opacity* instead of more widely used *blurring* for the purpose of distinguishing foveal vs. peripheral vision. Blurring more accurately approximates the difference in human perceptual abilities between the fovea and the periphery. However, our goal is not to approximate these differences, but simply to co-ordinate mouse and gaze locations. Because opacity makes the underlying image harder to see, we have found in pilot experiments that it strongly encourages mouse motion close to the gaze position. It will be an important next step to quantify this tendency, by comparing mouse and gaze positions in a larger study, using an accurate eye tracker.

VI. Conclusion and Future Work

As laid out in this methodology paper, FocalVid could be used to record and analyze participants' mouse-contingent visual selection data which can be used in laboratory studies, but importantly, is likely to be highly advantageous for online studies with remote participants. In this article, we detailed the completed components of the presented system. It should be noted that the presented system is functional as it stands to record and conduct the needed analysis of the data. The immediate expansions would elevate FocalVid to a better-suited tool for other researchers, with its capability to produce meaningful simplified outputs that they then could use towards their research.

A limitation of our approach is that mouse movements are slower than eye movements, for a given degree of accuracy [40]. This difference will limit our approach with respect to rapidly changing scenes. There are also other limitations associated with the presented system which should be taken into consideration when using the FocalVid interface:

- Participants can have different mouse types (travel mouse vs. professional mouse), which affects cursor tracking capabilities. These different mouse types could lead to variability in points of interest tracking precision. This should be taken into consideration when using FocalVid.
- Participants can have different computer screens with different sizes, contrast ratios and dynamic range. Differences in screens lead to different visibility in the semi-transparent region. Calibration protocols might be needed to compensate for this.
- Participants can have computers with different processing power capabilities. Processing power minimum requirements should be in place for participants.

In addition to these system limitations, there is an innate variability associate with participant populations which should be taken into consideration when using the FocalVid interface:

- Participants have different hand-eye coordination abilities (e.g. older persons may have deteriorated hand-eye coordination compared to younger participants, and other participants may suffer from conditions that impair hand-eye coordination). This variability should be taken into consideration while using the FocalVid platform.
- Participant environmental setting cannot be controlled precisely in remote studies (some participants might be much more distracted due to environmental factors). As such, attention evaluation protocols should be put in place in conjunction with this system to assess participant attention and engagement in the task.

The system limitations could be mitigated by future system improvements such as inclusion of a screen contrast calibration process for the present system. Participant variability limitations could be mitigated by incorporation of specific experimental conditions and participant selection criteria while utilizing the FocalVid platform.

Immediate future works relating to this project could be classified into time stamping improvements, in-depth system verification and data analysis additions. The client-side components of the presented system are processing intensive which could lead to a possibility of lags between produced time-stamp and video playback. Work is underway to improve upon the time-stamping method such that in addition to the present time-stamp, frame count for viewed video frames is also recorded to the database so that lags could be identified and adjusted for in the data processing step.

As part of the immediate future steps relating to this project, experimental participant data would be gathered and analysed. The authors are planning on testing of the gathered data against the already labeled Hollywood2 dataset [41]. Regarding data processing, semantic segmentation [42]–[45], and spatiotemporal data clustering and visualization [46]–[48] could be explored. Long-term system verification could be in the context of diversification of testing case scenarios to assess broader system functionality.

Acknowledgment

This work was supported by University of Waterloo's Social and Intelligent Robotics Research Lab (SIRRL). This research was undertaken, in part, thanks to funding from the Canada 150 Research Chairs Program.

References

[1] M. Argyle and M. Cook, *Gaze and Mutual Gaze.* Cambridge U Press, 1976.

[2] M. Jording, A. Hartz, G. Bente, M. Schulte-Rüther, and K. Vogeley, "The "Social Gaze Space": A taxonomy for gaze-based communication in triadic interactions," *Frontiers in psychology*, vol. 9, p. 226, 2018.

[3] M. Schulte-Mecklenbeck, R. O. Murphy, and F. Hutzler, "Flashlight–Recording information acquisition online," *Computers in human behavior*, vol. 27, no. 5, pp. 1771–1782, 2011.

[4] N. W. Kim, Z. Bylinskii, M. A. Borkin, K. Z. Gajos, A. Oliva, F. Durand, *et al.*, "BubbleView: An interface for crowdsourcing image importance maps and tracking visual attention," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 24, no. 5, pp. 1–40, 2017.

[5] M. C. Chen, J. R. Anderson, and M. H. Sohn, "What can a mouse cursor tell us more? correlation of eye/mouse movements on web browsing," in *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '01, Seattle, Washington: Association for Computing Machinery, 2001, pp. 281–282. DOI: 10.1145/634067.634234.

[6] Q. Guo and E. Agichtein, "Towards predicting web searcher gaze position from mouse movements," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '10, Atlanta, Georgia, USA: Association for Computing Machinery, 2010, pp. 3601–3606. DOI: 10.1145/1753846.1754025.

[7] K. Rodden, X. Fu, A. Aula, and I. Spiro, "Eye-mouse coordination patterns on web search results pages," in *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '08, Florence, Italy: Association for Computing Machinery, 2008, pp. 2997–3002. DOI: 10.1145/1358628.1358797.

[8] J. Huang, R. White, and G. Buscher, "User see, user point: Gaze and cursor alignment in web search," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12, Austin, Texas, USA: Association for Computing Machinery, 2012, pp. 1341–1350. DOI: 10.1145/2207676.2208591.

[9] A. Al-Rahayfeh and M. Faezipour, "Eye tracking and head movement detection: A state-of-art survey," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 1, pp. 2 100 212–2 100 212, 2013.

[10] D. Lagun and E. Agichtein, "Viewser: Enabling large-scale remote user studies of web search examination and interaction," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '11, Beijing, China: Association for Computing Machinery, 2011, pp. 365–374. DOI: 10.1145/2009916.2009967.

[11] S. Treue, "Visual attention: The where, what, how and why of saliency," *Current opinion in neurobiology*, vol. 13, no. 4, pp. 428–432, 2003.

[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[13] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[14] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, MIT press, 2006, pp. 155–162.

[15] A. Borji, "Saliency Prediction in the Deep Learning Era: An Empirical Investigation," *arXiv:1810.03716 [cs]*, Oct. 2018.

[16] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2015, pp. 1072–1080.

[17] A. R. Jansen, A. F. Blackwell, and K. Marriott, "A tool for tracking visual attention: The restricted focus viewer," *Behavior research methods, instruments, & computers*, vol. 35, no. 1, pp. 57–69, 2003.

[18] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011, ISBN: 0-19-162542-6.

[19] J. R. Bergstrom and A. Schall, *Eye Tracking in User Experience Design*. Elsevier, 2014, ISBN: 0-12-416709-8.

[20] Alfred L. Yarbus, *Eye Movements and Vision*. Plenum Press, 1967.

[21] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," in *Advances in Neural Information Processing Systems*, 1994, pp. 753–760.

[22] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, "Turkergaze: Crowdsourcing saliency with webcam based eye tracking," *arXiv preprint arXiv:1504.06755*, 2015.

[23] C. Shen, X. Huang, and Q. Zhao, "Predicting eye fixations on webpage with an ensemble of early features and high-level representations from deep network," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2084–2093, 2015.

[24] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2012.

[25] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th international conference on computer vision*, IEEE, 2009, pp. 2106–2113, ISBN: 1-4244-4420-9.

[26] W. Kienzle, F. A. Wichmann, M. O. Franz, and B. Schölkopf, "A nonparametric approach to bottom-up visual saliency," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds., MIT Press, 2007, pp. 689–696.

[27] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye

fixations," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4446–4456, 2017.

[28] J. Pan, E. Sayrol, X. Giro-i-Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 598–606.

[29] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, *et al.*, "Salgan: Visual saliency prediction with generative adversarial networks," *arXiv preprint arXiv:1701.01081*, 2017.

[30] F. Gosselin and P. G. Schyns, "Bubbles: A technique to reveal the use of information in recognition tasks," *Vision research*, vol. 41, no. 17, pp. 2261–2271, 2001.

[31] J. M. Henderson, K. K. McClure, S. Pierce, and G. Schrock, "Object identification without foveal vision: Evidence from an artificial scotoma paradigm," *Perception & Psychophysics*, vol. 59, no. 3, pp. 323–346, 1997.

[32] R. Bednarik and M. Tukiainen, "Validating the restricted focus viewer: A study using eye-movement tracking," *Behavior research methods*, vol. 39, no. 2, pp. 274–282, 2007.

[33] P. Tarasewich, M. Pomplun, S. Fillion, and D. Broberg, "The enhanced restricted focus viewer," *International Journal of Human-Computer Interaction*, vol. 19, no. 1, pp. 35–54, 2005.

[34] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2013, pp. 580–587.

[35] L. C. L. Silveira and V. H. Perry, "The topography of magnocellular projecting ganglion cells (M-ganglion cells) in the primate retina," *Neuroscience*, vol. 40, no. 1, pp. 217–237, 1991, ISSN: 03064522. DOI: 10.1016/0306-4522(91)90186-R.

[36] L. Chatriot, *Nedb*, https://github.com/louischatriot/nedb, 2018. (visited on 10/25/2020).

[37] E. M. Hoffman, S. Traversaro, A. Rocchi, M. Ferrati, A. Settimi, F. Romano, *et al.*, "Yarp based plugins for gazebo simulator," in *International Workshop on Modelling and Simulation for Autonomous Systems*, Springer, 2014, pp. 333–346.

[38] A. Pryke, S. Mostaghim, and A. Nazemi, "Heatmap visualization of population based multi objective algorithms," in *International Conference on Evolutionary Multi-Criterion Optimization*, Springer, 2007, pp. 361–375.

[39] M. Waskom, *Nedb*, https://github.com/mwaskom/seaborn, 2020. (visited on 10/25/2020).

[40] L. E. Sibert and R. J. K. Jacob, "Evaluation of eye gaze interaction," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '00, The Hague, The Netherlands: Association for Computing Machinery, 2000, pp. 281–288, ISBN: 1581132166. DOI: 10.1145/332040.332445.

[41] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2018, pp. 4894–4903.

[42] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[43] T. Zhu and D. Oved, *BodyPix - Person Segmentation in the Browser*, https://github.com/tensorflow/tfjs-models/tree/master/body-pix, 2020. (visited on 10/25/2020).

[44] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, 2018.

[45] Y. Li, J. Shi, and D. Lin, "Low-latency video semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2018, pp. 5997–6005.

[46] X. Li, A. Çöltekin, and M.-J. Kraak, "Visual exploration of eye movement data using the space-time-cube," in *International Conference on Geographic Information Science*, Springer, 2010, pp. 295–309.

[47] K. Kurzhals and D. Weiskopf, "Space-time visual analytics of eye-tracking data for dynamic stimuli," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2129–2138, 2013.

[48] K. Kurzhals, F. Heimerl, and D. Weiskopf, "Iseecube: Visual analysis of gaze data for video," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA '14, Safety Harbor, Florida: Association for Computing Machinery, 2014, pp. 43–50. DOI: 10.1145/2578153.2578158.