The Automated Temporal Analysis of Gaze Following in a Visual Tracking Task

The Automated Temporal Analysis of Gaze Following in a Visual Tracking Task

Vidushani Dhanawansa[1], Pradeepa Samarasinghe[1], Bryan Gardiner[2], Pratheepan Yogarajah[2], and Anuradha Karunasena[1]

[1] Faculty of Computing, Sri Lanka Institute of Information Technology, Sri Lanka
vidushani.d@gmail.com, {pradeepa.s, anuradha.s}@sliit.lk
[2] School of Computing, Engineering and Intelligent Systems, Ulster University, United Kingdom {b.gardiner, p.yogarajah}@ulster.ac.uk

Abstract. The attention assessment of an individual in following the motion of a target object provides valuable insights into understanding one's behavioural patterns in cognitive disorders including Autism Spectrum Disorder (ASD). Existing frameworks often require dedicated devices for gaze capture, focus on stationary target objects, or restrict the analysis to a single image obstructing a temporal analysis of the participant's response. This drives the motivation to address the persisting research gap in the analysis of the video capture of a visual tracking task of a dynamic physical object, over a time frame. To this end, this paper proposes a novel framework to analyse the temporal relationship between the 3D head pose angles and object displacement, and demonstrates its validity via application on the EYEDIAP video dataset. The conducted multivariate time-series analysis is two-fold; the statistical correlation computes the similarity between the time series as an overall measure of attention; and the Dynamic Time Warping (DTW) algorithm aligns the sequences of the target trajectory and the gaze, and computes relevant temporal metrics. It was proven that the correlation and metrics relevant to DTW corroborated each other. Additionally, the latency and maximum time of focus retention extracted as temporal characteristics, enabled an intragroup comparison between the performance of the participants. An extension to the analysis of the behavioural response concluded that the response to horizontal motion generally outperformed that of vertical motion, and that 56.3% of participants improved their retention of focus on the vertical motion of the target over time, implying that following a vertical target initially proved a challenging task.

Keywords: Automated gaze analysis · Multivariate time-series analysis · Head Pose.

1 Introduction

The analysis of human gaze with respect to a dynamic target is vital to the perception of one's attention and engagement as per clinical documentation [7]. In contrast to manual qualitative evaluation, merits of automated analysis include an unbiased judgement of the resultant gaze, enhanced by metrics to allow

further prediction based on the level of attention. Frameworks for analysis of this calibre have been applied in a variety of domains ranging from social behaviour of adults [8, 20, 21] and infants [13], classroom environments [4], [3] to human-robot interaction [17] and industrial environments [6].

Whilst gaze analysis on static images have been attempted with the aid of the line of sight [4] and heat maps [3,17] between the participant and target objects, only limited challenges in analysing a sequence of frames have been previously tackled. Lemaignan et al. [12] studied one's focus time on a target, determined by a broad field of attention estimated by the head pose, limiting the accurate alignment of the gaze to a small dynamic target. A study of joint attention [20] estimates the latency between the instruction and the resulting look, based on the presence of the gaze dot within the expected field of attention, whilst the latency, longest and shortest look for each object of interest has been previously assessed [21]. However, these work [20, 21] rely on specialised sensors to capture the gaze and object position, as opposed to a single video feed.

To date, computer vision-based eye gaze estimation remains challenging due to variations in head pose, illumination, occlusions and the requirement for the capture of high-resolution images of the eye [10]. Thus, the input modality of a majority of work in this domain constitutes of dedicated devices to facilitate eye tracking, which deviates from a natural interaction and contributes towards one's discomfort and expense of data collection. The aforementioned impracticality in eye tracking within a resource-restricted environment has recently expedited the approach of gaze estimation via head pose. Prior work has extensively proven the feasibility of robust head pose estimation in a wild environment from 2D images [16]. In addition, the reliability of this gaze estimation has been adequately demonstrated by experiments [18] which concluded that the head pose contributes to an average of 88.7% to one's focus of attention. Therefore, the model proposed in this paper focuses on exploiting the relationship between the 3D head pose angles and position of the object within a visual tracking task.

Furthermore, the analysis of attention over a time frame in response to a dynamic physical target is a persisting research gap in the domain of attention analysis, given the invalidity of prior approaches primarily focusing on stationary targets [12,20,21]. As per the Diagnostic and Statistical Manual of Mental Disorders [2], children displaying prodromal symptoms of Autism Spectrum Disorder (ASD) show deficits in following another's pointing, eye gaze or movement of a target object. For instance, the Autism Observation Scale for Infants (AOSI) [5],

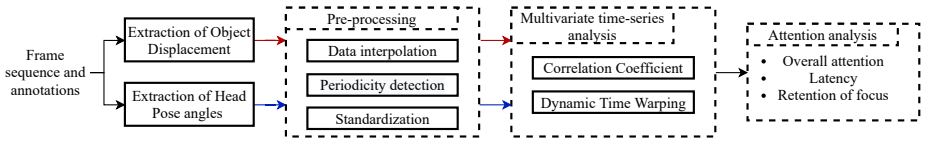


Fig. 1. System overview diagram of the proposed framework for the analysis of an object following task.

a standard set of semi-structured activities designed to detect and monitor early signs of autism, incorporates a visual tracking task which assesses the ability to track the trajectory of a rattle. The AOSI scoring system considers the smoothness of tracking the target object, the resultant delay in response, and ability to track the vertical movement of an object; factors which would be quantitatively assessed in the proposed approach.

Thus, an automated behavioural analysis of response to a dynamic target enhances the identification of symptoms of attention based disorders, and subsequent intervention to improve cognitive behaviour in children. To the best of the authors' knowledge, only a qualitative evaluation of a visual tracking task has been previously attempted [11], which visualises the variation of the estimated yaw to evaluate the smoothness of the participant's response. Furthermore, it fails to analyse the similarity between the head pose and the trajectory of the target object to assess how closely the target is followed, does not model the response to the vertical motion of the target, and is void of a temporal analysis of the response.

To address the aforementioned research gaps in the evaluation of an individual's engagement within a captured video of an object following task, the framework presented in Fig. 1 is proposed. The time-series data pertaining to the gaze and position of the target is subject to a multivariate time-series analysis, which employs statistical correlation to compute the similarity between the series, and the Dynamic Time Warping (DTW) algorithm [14] to further analyse the temporal dynamics. The DTW algorithm was applied to compute the time lag, and therefore parameters vital to draw a conclusion on the attention extended by an individual, in contrary to the application of DTW as an exclusive similarity measure between time series in domains including gesture recognition [1] and gait analysis [19]. The attention in response to the motion of the stimulus along the horizontal and vertical planes is based upon the relationship between the yaw and x displacement; and pitch and y displacement.

The potential for application of the framework in a clinical behavioural study of a group within a low-resource environment is demonstrated via the study on the EYEDIAP dataset [9]. To the best of our knowledge, we are the first to propose a simple computational framework of this calibre, with the following contributions:

- A multivariate time-series analysis between the head pose angles and object displacement for periodic motion, based on statistical correlation and Dynamic Time Warping.
- An intragroup analysis of the retention of focus and latency of gaze following, based on the deviation between the resulting and expected warping paths.
- An intragroup study of the behavioral response in following a visual target horizontally and vertically.

The rest of the paper is organized as follows: Sections 2 and 3 detail the methodology followed in the proposed framework. The analysed results and discussion is presented in Section 4. Finally, the paper is concluded in Section 5 with directions for future work.

2 Extraction of data and Pre-processing

2.1 EYEDIAP Dataset

The EYEDIAP video dataset, to the best of the authors' knowledge, is the only available annotated video dataset which captures participants gazing at a dynamic visual target (Fig. 2). The sessions of participants following a 3-D floating target (FT), while performing head movements (denoted as Mobile case, M) were extracted. The selected video input consisted of a standard RGB stream at resolution 640×480 and frame rate of 30 fps. The dataset comprises of 16 participants yielding a total of 18 videos as a result of participants 15 and 16 being recorded under two conditions. Video of participant 4 was discarded since 60.9% of frames lacked annotations.

2.2 Extraction of gaze and object displacement data

The fine-grained head pose of an individual is interpreted as a continuous angular measurement across multiple Degrees of Freedom (DOF) [15]. The interpretation of 3 DOF is adopted in this work, where,motion about a vertical, horizontal and longitudinal axis, is denoted by the change in yaw(θ), pitch(ψ) and roll(ϕ), respectively. Therefore, a proportional change in yaw and pitch is expected in response to the motion of the target along the horizontal and vertical planes.

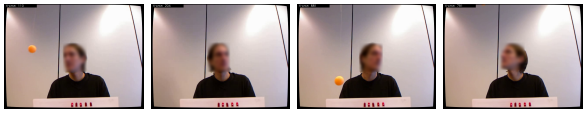
The dataset comprises of gaze annotations in the form of a 3D rotational matrix for each frame. The relationship between the R and the gaze angles is defined by (1) which was manipulated to extract the head pose angles of yaw and pitch. Furthermore, the coordinates of the spatial center of the ball along the horizontal (x) and vertical (y) axes for each frame, were extracted as a representation of the displacement.

$$R = \begin{bmatrix} \cos\psi\cos\phi & \begin{matrix}\cos\psi\sin\theta\sin\phi \\ -\sin\psi\cos\phi\end{matrix} & \begin{matrix}\cos\psi\sin\theta\cos\phi \\ +\sin\psi\sin\theta\end{matrix} \\ \sin\psi\cos\phi & \begin{matrix}\sin\psi\sin\theta\sin\phi \\ +\cos\psi\cos\phi\end{matrix} & \begin{matrix}\sin\psi\sin\theta\cos\phi \\ -\cos\psi\sin\theta\end{matrix} \\ -\sin\theta & \cos\theta\sin\phi & \cos\theta\cos\phi \end{bmatrix} \tag{1}$$

A proportional relationship between the pitch and y displacement, and inversely proportional relationship between the yaw and x displacement was evident as in Fig. 3(a), (b)-1, (b)-2. Thus, an inverted version of the yaw was considered for further analysis.

2.3 Pre-processing

Data interpolation In addition to instances where the target exceeded the frame of view (Fig. 2(b), (d)), the displacement lacked a significant quantity of annotations. On average, 20% of frames per video lacks annotations as a



(a) (b) (c) (d)

Fig. 2. Sample frames of video 10, with images modified to maintain privacy of participants: (a), (b) Following the horizontal motion of the target; (c), (d) Following the vertical motion of the target.

result. Therefore, available data was interpolated as seen in Fig. 3(b)-1, (b)-2 to generate the x and y coordinates of the missing segments taking into account the gradients of the adjoining segments which represent the velocity of the target.

For a segment between $[t_1, t_2]$, midpoint $\bar{t}$, and gradients of adjoining segments m_1 and m_2, the interpolated displacement value $d_{int}[t]$ at a time instant t was generated by (2). For additional lacking annotations within the linear regions of displacement, linear interpolation was applied via (3).

$$d_{int}[t] = \begin{cases} m_1 - \frac{m_1.(t-t_1)}{\bar{t}-t_1} + d[t-1], & \text{if } t_1 < t < \bar{t}. \\ \frac{m_2.(t-\bar{t})}{t_2-\bar{t}} + d[t-1], & \text{if } \bar{t} \leq t < t_2. \end{cases} \tag{2}$$

$$d_{int}[t] = d[t_1](1-\mu) + d[t_2]\mu \tag{3}$$

where $\mu = \frac{t-t_1}{t_2-t_1}$. Segments exceeding 90 frames were not interpolated since the approximations made over a larger number of frames may be compromised.

Periodicity detection and segmentation In order to analyse the attention of an individual over periodic motion of the target, the segments of strictly periodic motion along the horizontal or vertical axis were identified (Fig. 3(c)-1, (c)-2). All local maxima were identified by simple comparison with the neighboring values. The minimal horizontal distance between samples was restricted to 75, and the prominence was limited to 200 to eliminate minor peaks. A threshold of 360 frames was applied in identifying the consecutive peaks, since it was evident that the period of cycles of motion in the video dataset did not exceed this value.

Standardization Given that the gaze angles and displacement lie within dissimilar ranges, all data segments were standardized by (4) where μ and σ denote the mean and standard deviation of the segment, respectively.

$$x_{std} = \frac{x - \mu}{\sigma} \tag{4}$$

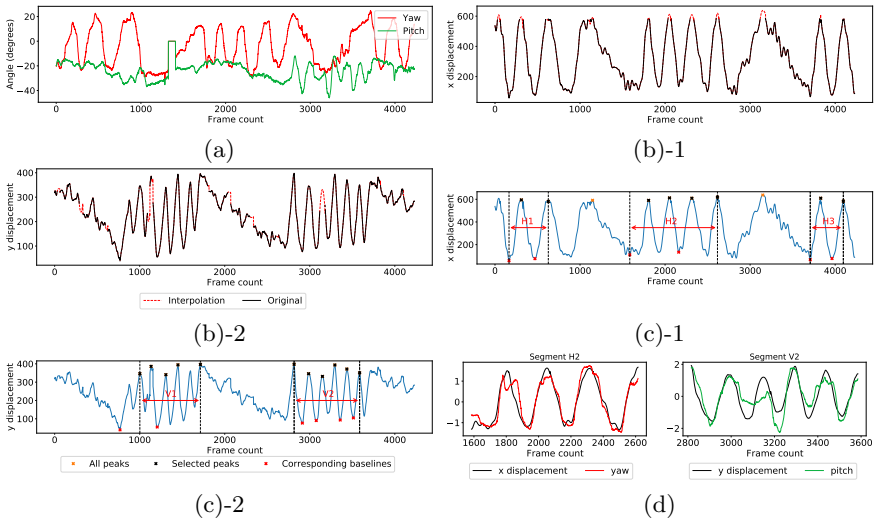


Fig. 3. Extracted data and pre-processing for video 9: (a) Extracted yaw and pitch angles; (b) Original and interpolated displacement; (c) Automatic periodicity detection and segmentation of displacement. Segments of horizontal and vertical motion are denoted as H1-H3, and V1-V2, respectively.; (d) Relationship between standardized data for selected segments H2 and V2 from (c).

3 Time-series analysis between the gaze and object displacement

3.1 Statistical correlation

The Pearson's correlation coefficient r, computed by (5) evaluates the linear relationship between two continuous variables, x and y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2)} \tag{5}$$

The coefficients between the yaw and x displacement, and pitch and y displacement, evaluated for each identified segment, served as an overall measure of the attention extended by the participant over the considered time frame. The coefficients for segments H2 and V2 in Fig. 3(d) are 0.941 and 0.816, respectively.

3.2 Dynamic Time Warping (DTW)

The Dynamic Time Warping (DTW) algorithm yields the optimal alignment between two non-linearly warped time-series by computing the corresponding points bearing the least cost between the two series. Thus, it facilitates further analysis of the time lag between the gaze and trajectory of the target.

Implementation Let the two sequences be defined as $X = (x_1, ..., x_n)$ and $Y = (y_1, ..., y_m)$. The generated cost matrix $d_{n \times m}$ is computed as $d_{i,j} = |x_i - y_j|$ where i and j refer to the indices on the sequences X and Y, respectively. The warping path, W of length k, is defined as the sequence of points of minimum overall cost and is represented as:

$$W = (w_1, ..w_t, .., w_k), \quad w_t = (i_t, j_t) \tag{6}$$

where w_t refers to a point lying on the warping path, W at time, t. The following restrictions apply in generating the warping path as in Fig. 4(a):

- Boundary condition: The path begins and ends with the starting and ending points of both signals, respectively, such that $w_1 = (1, 1)$, $w_k = (n, m)$.
- Monotonicity condition: The time order is preserved such that $i_{t-1} \leq i_t$ and $j_{t-1} \leq j_t$.
- Continuity condition: The translation of the path is restricted to adjacent points in the matrix. Therefore, $i_t - i_{t-1} \leq 1$ and $j_t - j_{t-1} \leq 1$.
- Warping window constraint: The window length (w), the maximum deviation of the warping path from the diagonal is restricted such that $|i - j| \leq w$. Based on empirical evidence, a window of 150 was applied.

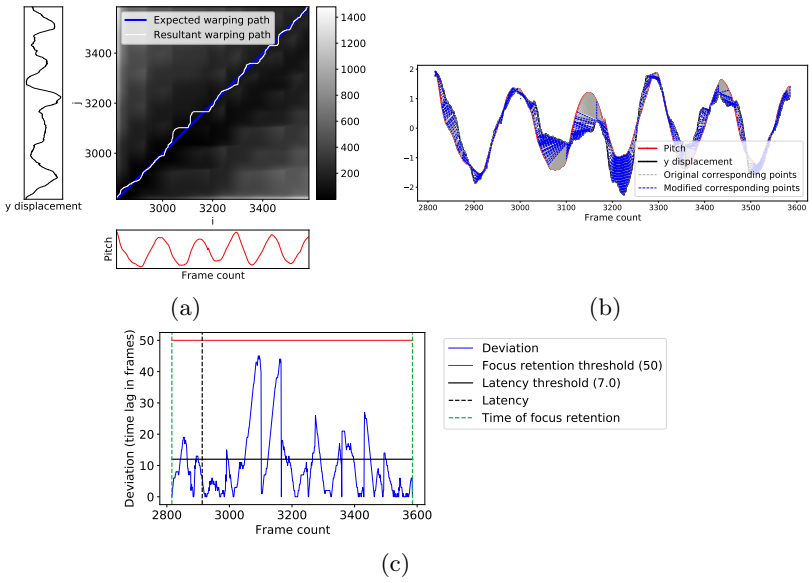


Fig. 4. Results of the DTW algorithm for segment V2 of video 9: (a) Generated cost matrix, expected warping path for perfect mapping, and the resultant warping path of minimum overall cost; (b) Original and modified alignment between the gaze and displacement sequences; (c) Time lag analysis of focus retention time and latency.

Analysis of DTW metrics The accumulated cost of the warping path is a measure of the association between the two sequences aligned by the DTW approach. Therefore, a metric, Normalised Accumulated Cost (NAC) was defined as the derived cost normalised by the length of the warping path to account for sequences of varied lengths. Based on the continuity condition, the warping path includes all indices of each sequence, implying the repetition of indices and a path of variable length which hinders an accurate temporal analysis. Therefore, considering the displacement as the baseline, a modified warping path included only the corresponding index of minimum cost for each index of the displacement, as visualised in Fig. 4(b). Thereafter, the difference between the corresponding indices, normalised by the length of the warping path is calculated, which is interpreted as the time lag. The retention of focus and latency is evaluated (Fig. 4(c)) based on this metric termed as Normalised Time Deviation (NTD).

The **time of focus retention** refers to the maximum duration an individual maintains a reasonable level of focus on the target object. Therefore, this was obtained by identifying the maximum number of consecutive frames of time lag less than a threshold of 45 frames, or 1.5s within this framework. In addition, the percentage of focus retention over the number of frames of the segment was computed. This is similar in concept to the normalized ratio of time that the human interactant focuses its attention on the target, denoted as 'with-me-ness' in Lemaignan et al. [12]. The **latency** is the time elapsed prior to the gaze of the individual aligning closely with the path of the target object with certain accuracy, during which the time lag, or deviation should remain 0 over a time period of at least 1.5s. However, a threshold equal to the 1^{st} quartile of the time lag was considered to account for minor fluctuations. Interpreting Fig. 4(c), the individual retained the focus for the complete duration of 770 frames, i.e. 25.6s. Furthermore, a latency of 3.23s was recorded given the that 97 frames elapsed prior to the target being accurately followed. While the latency was computed for each segment, the overall latency of a participant was considered as that of the initial segment of the task.

4 Results and Discussion

A detailed study of the results obtained via the application of the framework on the EYEDIAP dataset is presented in this section. The study is further extended to comment on the discernible behavioural characteristics of the participants.

4.1 Multivariate time-series analysis metrics

The metrics relevant to correlation and the DTW algorithm were compared to ensure the validity of the chosen approaches for the multivariate time-series analysis. It was hypothesised that an inverse relationship would result between the correlation coefficient, and the DTW metrics of NAC and NTD since a well-matched head pose series to the displacement yields a high correlation, and a minimum NAC and NTD. The scatter plot in Fig. 5 which includes metrics of all

extracted segments in the dataset depicts an inversely proportional relationship between the relevant metrics, validating our hypothesis.

While it is understood that the two selected approaches of time-series analysis within this application corroborate each other, the significance of the approaches lie in their interpretation. The correlation coefficient, a single metric in the range [-1, +1] is an overall measure of the similarity between the head pose and the position of the target. The NAC is an enhanced similarity measure yielding the average cost between the sequences following the alignment by the DTW approach, and denotes how closely the trajectory of the head pose aligns with that of the target, removing dependancy on time. In contrast, the NTD is the average time lag between the aligned sequences, serving as a temporal measure between the time-series.

4.2 Intragroup comparison of temporal characteristics

A comparison between the performance of the participants in terms of the temporal features was conducted to further analyse the participants' level of attention. This considered the maximum percentage of focus retention amongst all segments within the video, and the latency of the initial segment as in Fig. 6. The focus retention measures the maximum time a participant is capable of maintaining focus on the target, with a minimal lag. It was evident that participants 1, 5, 8, 12 and 16_B were unable of maintaining focus for more than 50% of a segment. In contrast, the latency is a strict measure of the time taken for the gaze of the participant to align with the expected trajectory of the target. The latency of the initial segment was utilised to extract the behaviour prior to any training the participant may accumulate during the task. Here, participants 1, 2, 5, 10, 15_B recorded comparably high latency values denoting that a significant time elapses prior to the gaze strictly aligning with the target position.

Taking into account both these factors, participants 1 and 5 were less successful in gaze following, bearing a low retention of focus and relatively high latency. Participants 6, 11 and 16_A showed 100% of focus retention and negligible latency depicting excellent performance. Participant 10 showed an interesting outcome with a focus retention of 100% and the maximum latency of the group, implying the maintenance of moderate focus throughout while failing to strictly follow the target.

An analysis of the latency and focus retention time of all segments revealed that in 58.8% of segments, the latency occurred before the focus retention time, signifying that individuals are more likely to retain focus continuously after having reached the expected trajectory even for a short period of time. In contrary, the latency occurred within the focus retention time in 29.4% of segments. This implies that certain individuals needed to focus on the target for a certain period of time prior to reaching the expected trajectory. The latency was not recorded in 11.8% segments indicating that participants failed to closely follow the target.

A comparison of the percentage of focus retention between the initial and final segments along the horizontal and vertical planes saw an increase in focus retention for vertical motion in 56.1% of participants. This notable improvement

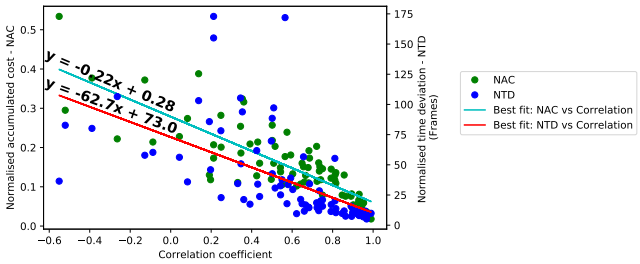


Fig. 5. Inverse relationship between Correlation Coefficient and DTW metrics (i.e. Normalized Accumulated Cost and Normalised Time Deviation) of all identified segments in the dataset.

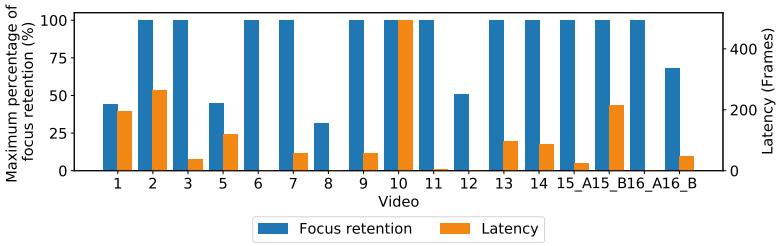


Fig. 6. Comparison of the maximum recorded focus retention percentage and the latency between participants.

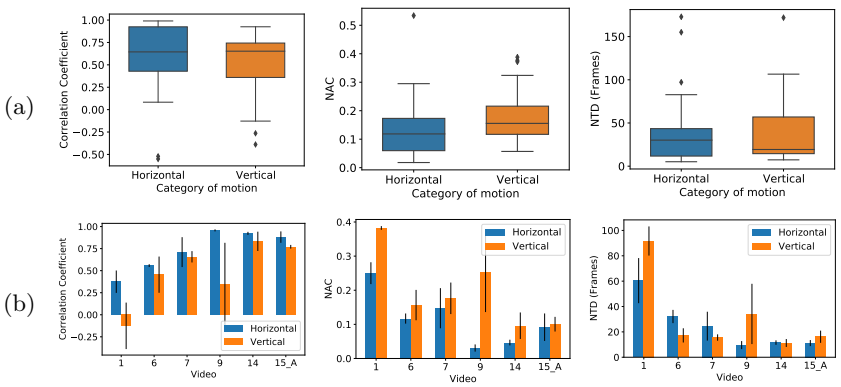


Fig. 7. Variation of Correlation Coefficient and DTW metrics of Normalised Accumulated Cost (NAC) and Normalised Time Deviation (NTD) in response to horizontal and vertical motion. (a) Box plots of all segments in the dataset; (b) Mean and standard deviation (represented by error bars) of all segments in sample videos. Selection of videos exhibit the possible variations in participant response within the dataset.

may imply that participants were involuntarily trained to focus on the target object over the time frame of the task. In contrast, a 42.9% of participants showed a declining focus retention percentage for horizontal motion.

4.3 Intragroup comparison of response to horizontal and vertical motion

A study of the association between the metrics and the plane of motion of the target (Fig. 7) indicated that in general, participants were more responsive towards motion along a horizontal plane. Horizontal motion has resulted in a notably higher correlation coefficient and lower NAC throughout the dataset observing the 1^{st} and 3^{rd} quartiles in Fig. 7(a). The results of sample videos in Fig. 7(b) confirm the same. This observation, along with the improvement in focus retention by the completion of the task may indicate that the vertical motion of the stimulus proves challenging, leading to reduced attention. In contrast, the declining or relatively constant focus could be justified by the monotonous nature of the horizontal motion, given that the object lies upon a constant level of sight.

It is also worth noting that the NTD metric shows a less significant bias towards horizontal motion as evident through the median value of the boxplot figures. This is further highlighted by the sample results which depict that horizontal motion in videos 6 and 7 has yielded a higher NTD. This occurs since this metric emphasizes completely on the resulting time gap between the aligned sequences, whereas the correlation and NAC measures the similarity between the sequences. This suggests that the gaze of participants 6 and 7 follow the horizontal trajectory of the target, albeit with a significant lag.

5 Conclusion and Future Works

This paper proposes a novel framework for the automated evaluation of parameters pertaining to one's gaze response to the motion of a physical target, addressing the research gaps in the lack of an approach for gaze analysis focusing on a dynamic physical target, conducting a temporal analysis of the performance and void of wearables for gaze capture. The time-series analysis between the head pose and the object displacement resulted in the correlation coefficient as an overall similarity measure. The analysis was enhanced by the similarity measure following the alignment of the time-series via the DTW algorithm, and the resultant time gap, upon which the time of focus retention and latency was computed. The application of the algorithm to the EYEDIAP video dataset showed that the correlation and the metrics of DTW corroborated each other. Based on the temporal features, the gaze responses of participants were compared. While an increasing trend in focus retention between the initial and final segments of vertical motion was observed amongst participants, the response towards horizontal motion of the target outperformed that of vertical motion, as evident through all metrics.

Given the necessity of only a single video capture, the proposed framework shows prospects for implementation in low-resource environments such as ASD screening centers for children in developing countries. Thus, future work includes the application of the model on collected clinical data to evaluate the gaze response of children in the identification of deficits pertaining to ASD. A deep learning model for head pose estimation and a combination of object detection and tracking algorithms would replace the phase of extracting dataset annotations of gaze and displacement as done in this study. Furthermore, the proposed algorithm would incorporate an alternative framework based on the analysis of eye gaze angles, facilitating a comparison between the two modes for gaze analysis. Finally, the development of a model for the automated categorisation of a child's gaze response according to the AOSI standard, based on the computed correlation and temporal metrics, is worth investigating.

Acknowledgment

This research was supported by the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education of Sri Lanka funded by the World Bank (`https://ahead.lk/result-area-3/`).

References

1. Alon, J., Athitsos, V., Yuan, Q., Sclaroff, S.: A unified framework for gesture recognition and spatiotemporal gesture segmentation. IEEE transactions on pattern analysis and machine intelligence **31**(9), 1685–1699 (2008)
2. Association, A.P., Association, A.P., et al.: Diagnostic and statistical manual of mental disorders: Dsm-5. Arlington, VA (2013)
3. Aung, A.M., Ramakrishnan, A., Whitehill, J.R.: Who are they looking at? automatic eye gaze following for classroom observation video analysis. International Educational Data Mining Society (2018)
4. Bidwell, J., Fuchs, H.: Classroom analytics: Measuring student engagement with automated gaze tracking. Behav Res Methods **49**, 113 (2011)
5. Bryson, S.E., Zwaigenbaum, L., McDermott, C., Rombough, V., Brian, J.: The autism observation scale for infants: scale development and reliability data. Journal of autism and developmental disorders **38**(4), 731–738 (2008)
6. Cenzi, E.D., Rudek, M.: A method to gaze following detection by computer vision applied to production environments. In: IFIP International Conference on Product Lifecycle Management. pp. 36–49. Springer (2020)
7. Emery, N.J.: The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience & biobehavioral reviews **24**(6), 581–604 (2000)
8. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1226–1233. IEEE (2012)
9. Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 255–258 (2014)

10. Ghosh, S., Dhall, A., Hayat, M., Knibbe, J., Ji, Q.: Automatic gaze analysis: A survey of deep learning based approaches. arXiv preprint arXiv:2108.05479 (2021)
11. Hashemi, J., Spina, T.V., Tepper, M., Esler, A., Morellas, V., Papanikolopoulos, N., Sapiro, G.: A computer vision approach for the assessment of autism-related behavioral markers. In: 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL). pp. 1–7. IEEE (2012)
12. Lemaignan, S., Garcia, F., Jacq, A., Dillenbourg, P.: From real-time attention assessment to "with-me-ness" in human-robot interaction. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 157–164. Ieee (2016)
13. Michel, C., Kayhan, E., Pauen, S., Hoehl, S.: Effects of reinforcement learning on gaze following of gaze and head direction in early infancy: An interactive eye-tracking study. Child Development (2021)
14. Müller, M.: Information retrieval for music and motion, vol. 2. Springer (2007)
15. Neto, E.N.A., Barreto, R.M., Duarte, R.M., Magalhaes, J.P., Bastos, C.A., Ren, T.I., Cavalcanti, G.D.: Real-time head pose estimation for mobile devices. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 467–474. Springer (2012)
16. Patacchiola, M., Cangelosi, A.: Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. Pattern Recognition **71**, 132–143 (2017)
17. Saran, A., Majumdar, S., Short, E.S., Thomaz, A., Niekum, S.: Human gaze following for human-robot interaction. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8615–8621. IEEE (2018)
18. Stiefelhagen, R.: Tracking focus of attention in meetings. In: Proceedings. Fourth IEEE International Conference on Multimodal Interfaces. pp. 273–280. IEEE (2002)
19. Veeraraghavan, A., Roy-Chowdhury, A.K., Chellappa, R.: Matching shape sequences in video with applications in human movement analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(12), 1896–1909 (2005)
20. Venuprasad, P., Dobhal, T., Paul, A., Nguyen, T.N., Gilman, A., Cosman, P., Chukoskie, L.: Characterizing joint attention behavior during real world interactions using automated object and gaze detection. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications. pp. 1–8 (2019)
21. Venuprasad, P., Xu, L., Huang, E., Gilman, A., Ph. D, L.C., Cosman, P.: Analyzing gaze behavior using object detection and unsupervised clustering. In: ACM Symposium on Eye Tracking Research and Applications. pp. 1–9 (2020)