Evaluating Calibration-free Webcam-based Eye Tracking for Gaze-based User Modeling

Stephen Hutt

University of Denver, USA, stephen.hutt@du.edu

Sidney K. D'Mello

University of Colorado Boulder, USA, Sidney.dmello@colorado.edu

Eye tracking has been a research tool for decades, providing insights into interactions, usability, and, more recently, gaze-enabled interfaces. Recent work has utilized consumer-grade and webcam-based eye tracking, but is limited by the need to repeatedly calibrate the tracker, which becomes cumbersome for use outside the lab. To address this limitation, we developed an unsupervised algorithm that maps gaze vectors from a webcam to fixation features used for user modeling, bypassing the need for screen-based gaze coordinates, which require a calibration process. We evaluated our approach using three datasets (N=377) encompassing different UIs (computerized reading, an Intelligent Tutoring System), environments (laboratory or the classroom), and a traditional gaze tracker used for comparison. Our research shows that webcam-based gaze features correlate moderately with eye-tracker-based features and can model user engagement and comprehension as accurately as the latter. We discuss applications for research and gaze-enabled user interfaces for long-term use in the wild.

CCS CONCEPTS •Human-centered computing~Human computer interaction (HCI)~HCI design and evaluation methods~User models**•Human-centered computing~Human computer interaction (HCI)**

Additional Keywords and Phrases: Eye Tracking, Webcam, Mind Wandering, Comprehension, DBSCAN, Gaze-enabled interfaces

1 INTRODUCTION

Our eyes provide a window into our minds [18, 25, 49]. Eye gaze is considered a real-time index of the information-processing priorities of the visual system because physiological and cognitive limitations on vision, attention, and memory require the eyes to shift from location to location (called gaze saccades) to construct a comprehensive representation of the external world. Thus, it is generally assumed that attention is focused on where the eyes are fixated (called gaze fixations, e.g., [25, 61]). Eye tracking is, therefore, a very attractive methodology for multimodal analysis and user modeling. Researchers have shown that knowing where a users' gaze is focused can be leveraged to model various aspects of cognition, emotion, and social dynamics [3, 11, 15, 24, 29, 35]. This information can be used to dynamically adjust the user's experience and create a more personalized, responsive interaction.

One caveat to incorporating eye tracking in multimodal work is the eye tracking device itself. Most traditional eye tracking research (e.g., see review [49]) has used research-grade eye trackers that cost tens of thousands of dollars and are not portable, thereby limiting scalability. Commercial, off-the-shelf (COTS) eye trackers offer a cheaper, portable, and scalable alternative to research-grade equipment, allowing eye tracking research to be conducted outside of the lab [26]. However, COTS eye trackers come with additional financial and setup costs and are difficult to seamlessly integrate within everyday computing environments.

As an alternative, researchers have been exploring webcam-based eye tracking (reviewed below), which entails deriving gaze coordinates without using an eye tracker [33, 43, 57]. An obvious advantage is that webcams are ubiquitous in modern computers or can be inexpensively purchased. However, one considerable barrier is that most existing

approaches require users' to undergo a calibration procedure, which is the process of converting data that is relative to the camera being used (i.e., in real-world coordinates) into data relative to the UI (i.e., screen coordinates) [63]. During the calibration process, the user's gaze is directed to certain points on the screen at certain times, thus training a conversion algorithm that converts the three-dimensional gaze vector to two-dimensional screen coordinates. This process can be time-intensive, in some cases requiring multiple attempts to train a stable conversion or needing to recalibrate if the users' position changes. Such a process is not suited for several gaze-enabled UIs, especially in public settings such as interactive displays or where time is limited.

Can we leverage the convenience of webcam-based eye tracking but without calibration? This paper proposes a novel technique to derive approximations of eye gaze features from the raw three-dimensional gaze vectors derived from webcam data without calibration and without incorporating any additional information (e.g., user position). The key insight is that many user-modeling problems do not require precise screen-based coordinates, instead utilizing higher-order gaze features derived from eye movements, including fixations and saccades. By bypassing the step of computing screen-based gaze-coordinates, we can create a subset of the gaze features without calibration, this is especially useful for post-hoc video analysis where calibration was not conducted, or for real-time applications where opportunities for calibration may be limited. We evaluate the method on three previously collected datasets, encompassing different interfaces (computerized reading or an intelligent tutoring system), environments (laboratory or classroom), and types of comparison gaze trackers.

1.1 Background and Related Work

A conventional eye tracker typically uses infrared-based techniques, commonly referred to as PCCR (pupil center and cornea reflection). Though this technique is the 'gold standard' for eye tracking, it requires specialized equipment (i.e., an eye tracker) to illuminate the eye with infrared light and record the reflection [23]. This reflection is then used to calculate the real-world gaze trajectory, which is then (via calibration) converted into screen coordinates. Such a specialised approach can be costly, with research-grade equipment costing in excess of 30,000 USD. However, the recent development of commercial off-the-shelf (COTS) eye trackers (retailing for hundreds of dollars rather than thousands) has enabled an exciting wave of "in the wild" eye tracking, for a fraction of the cost.

Cheaper still is the option to use data from traditional standard webcams (e.g., no infrared) to produce gaze estimates. Appearance- [2, 6, 60] and shape- [46, 50] based techniques can be used to generate gaze estimations from images captured by a webcam. Though cheaper and requiring no specialized equipment, these techniques have been shown to be less accurate than PCCR approaches [31, 33, 63], posing issues for some gaze enabled Uis.

To account for inaccuracies, prior studies have examined if eye gaze was focused on broad areas of interest (AOIs) instead of finer grain locations. This entails making explicit assumptions about where the user should be looking (e.g., at the camera [62] or the video window [38] for a video conference). This approach might be suitable for specific applications with well-defined AOIs, but not more generally. To address this limitation, Tran et al. [56] used unsupervised clustering to automatically detect AOIs based on where gaze was most frequently clustered. This approach is viable when clusters can be detected but not in other cases, such as reading. Further, they do not provide detail regarding key eye movements such as fixations or saccades, which are needed for several applications.

Recent work compared gaze calculated from a smartphone camera to a Tobii Glasses Pro 2 wearable eye tracker [57] while participants completed a series of tasks such as fixation tasks (looking at a certain point), visual search, and reading comprehension. The authors leveraged a convolutional neural network (CNN) initially trained on the public GazeCapture Dataset [35], with additional calibrations per user to refine the model. The authors found the results to be

comparable [57], both in terms of the number of fixations calculated and the location of gaze. Though promising, it is not known if eye gaze could be as accurately calculated if the data were uncalibrated. Similarly, all gaze computations from the video data were computed offline so as not to overtask the smartphone. It is not clear if this model could run as effectively in real-time.

There have been attempts to avoid calibration entirely. Using a PCCR tracker, [1] leveraged the gaze patterns of users with calibrations to auto-calibrate those without calibrations. This relies on a dataset where the same stimuli are used. Sugano et al. used a similar approach [55] but compared eye gaze to a saliency map of the stimuli. These approaches were successful, but both had a high error rate (3-5 visual degrees) and required detailed knowledge of the screen content. SearchGazer [42] avoided a formal calibration for webcam based tracking by using implicit calibration, which assumed that the user gazed at a button when it was clicked. The resultant data of clicks and gaze coordinates was used to regress screen- on webcam- gaze location. This approach had moderate success over time but requires knowing the location of UI elements and does not work when such interactions are infrequent (e.g., reading text).

1.2 Current Work, Novelty, and Contribution

We make two contributions. First, we present an unsupervised method based on the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm [52] to derive an approximation of fixations (a common metric for user modelling) from uncalibrated webcam data. Our proposed data-driven approach requires no additional detail about the users' screen or calibration information, potentially enhancing generalizability. Second, we evaluate our methodology on three diverse, previously collected datasets using two evaluation criteria: (a) correlation of webcam-based gaze features with those derived from a traditional PCCR gaze tracker; (b) evaluating the accuracy of using webcam-based gaze features with PCCR-based features for gaze-based user modelling.

The datasets themselves had several differences, allowing us to test the robustness of the approach. First, we consider how webcam gaze estimations compare to a research-grade eye tracker to model mind wandering (an attentional shift from task-related processing to internal task-unrelated thoughts [53]) while users completed a computerized reading task in the laboratory (Dataset 1). This controlled environment, combined with a high fidelity eye tracker and predictable eye movements (based upon decades of reading research), allows us to evaluate the initial feasibility of our approach. For our next dataset (Dataset 2), we also considered webcam gaze estimations in the lab, but instead focused on modeling reading comprehension rather than mind wandering. It served as a conceptual replication of Dataset 1. Our third dataset also focused on modeling mind wandering, but in a more authentic "in-the-wild" environment, specifically, a high school classroom (Dataset 3) and with a multimedia learning technology [41]. We also used a COTS PCCR eye tracker, rather than a research-grade, eye tracker for comparison in Dataset 3.

To our knowledge, this is the first study to attempt to use unsupervised fixation estimations from uncalibrated gaze vectors estimated from webcam data for user modelling tasks. This work serves as an additional step towards moving decades of eye tracking research into ecologically valid environments, especially those where regular calibration may be undesirable or impossible.

2 FIXATION APPROXIMATION ALGORITHM

Our high-level pipeline is shown in Figure 1. We used the open-source software OpenFace [2, 4] to obtain camera-centered gaze vectors (x, y, and z coordinates for each eye) for each video frame in which a face has been detected. OpenFace uses a Constrained Local Neural Field landmark detector [4, 59] to detect eyelids, iris, and the pupil trained on the SynthesEyes training dataset [59]. It calculates the 3D camera coordinates for each pupil and the eyeball center;

these two points are then used to calculate an estimated gaze vector, a three vector coordinate relative to the camera's position. For example, if a user were looking directly into the camera, the vector would be [0,0,0]. Although we use OpenFace here (due to availability of existing data), the approach described should work with an alternate system that extracts gaze estimates from face meshes.

Next, we use unsupervised clustering to approximate gaze fixations from the camera-centered gaze vectors. We adapt the DBSCAN [19, 52] algorithm as it does not require a pre-specified number of clusters, meaning we don't need to know the number of fixations before we start. The standard DBSCAN algorithm clusters on one dimension and has been used for unsupervised AOI detection [56]. However, to be considered part of a fixation, samples must be close temporally as well as spatially, requiring us to cluster on two dimensions.

The updated DBSCAN algorithm has three hyperparameters. The minimum number of samples required to form a new fixation, the spatial threshold (the maximum radius of a fixation), and the temporal threshold (the temporal search radius around a point). The algorithm iterates through each point in the data that has yet to be examined. It makes a copy of the data and filters by the temporal threshold such that items that remain are within the temporal range of the current point. If the number of points within the temporal distance meets the criteria for a new fixation (e.g., more than the minimum number of samples), a new fixation is created, with the centroid is set to the current point. Next, the algorithm examines each of the data points in the filtered list in temporal order to see if they are within the spatial threshold (calculated as the Euclidian distance between the vectors) of the fixation centroid. If so, the sample is added to the fixation. The fixation centroid is recalculated, and additional samples within the temporal threshold are added to the filtered list. The temporal threshold thus scopes the clustering; however, as more samples are added to a fixation, the temporal window for that fixation also grows. The process repeats for all samples in the data, skipping over samples that have already been visited as a given sample may not be part of multiple fixations. This is then followed by a data cleaning step (described below). A simplified example (in 2D) for forming one fixation is shown in Figure 1. The pseudocode for the algorithm is shown in Figure 2.

Of the three hyperparameters, the spatial threshold is relative to the camera and user positioning and must be tuned. In contrast, the temporal threshold and the minimum number of samples can be set based upon the sampling rate of the video and known data about the average length of fixations [30, 49]. DBSCAN has big O complexity of O(nlog(n)), this is maintained in our adaptation.

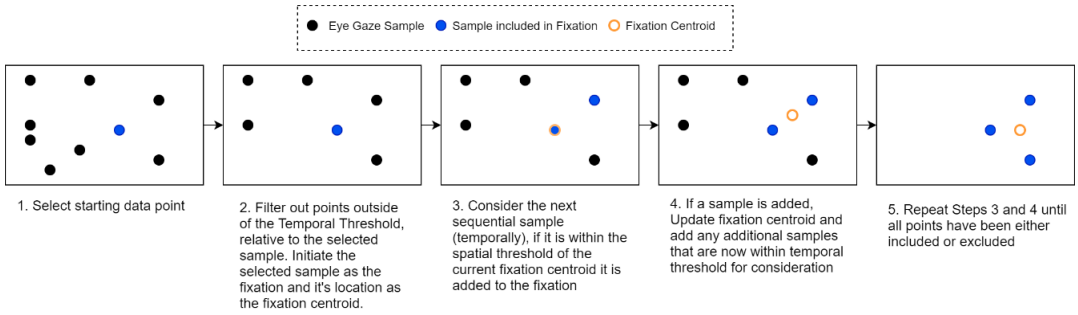

Figure 1. Simplified process (in 2D) for assigning samples to a fixation

```
Function CalculateFixations(Samples, SpatialThreshold, TemporalThreshold, MinimumSamples)
    for P in Samples do
        if P is visited then
            Continue
        end
        Mark P as visited
        neighbours ← getNeighbours(P, SpatialThreshold, TemporalThreshold)
        if neighbours.length < MinimumSamples then
            //Ignore P as Noise
        else
            C = newFixation
            expandCluster(P,C,neighbours SpatialThreshold, TemporalThreshold, MinimumSamples)
        end
    end
    Clean Fixations()

Function expandCluster(P,C, neighbours, SpatialThreshold, TemporalThreshold, MinimumSamples)
    add P to fixation C
    for P' in neighbours do
        Mark P' as visited
        add P' to fixation C
        Update Centroid of fixation C
        Neighbours' ← getNeighbours(C.centroid)
        if Neighbours'.length > MinimumSamples then
            neighbours = neighbours + Neighbours'
        end
    end

Function getNeighbours(P, SpatialThreshold, TemporalThreshold)
    Filter Data by TemporalThreshold
    Remove visited points Return points within SpatialThreshold
```

Figure 2. Pseudocode for Fixation Approximation Algorithm

Data Cleaning. The candidate fixations are then post-processed to address inaccuracies from the initial clustering. Specifically, the algorithm can assign consecutive samples to different fixations resulting in implausible situations, for example, three consecutive samples assigned to fixations, 1, 2, and 1, respectively. To address this, we remove the fixation assignment from any point that does not share a consecutive sample with that same cluster (Figure 3A). Following this step, a fixation may be separated by an unassigned sample. In this case, we split the fixation into two separate fixations (providing that both satisfy the criterion as specified by the algorithm hyperparameters). If a fixation does not include enough samples after splitting, it is removed (see Figure 3B).

Finally, we check the fixation duration. Given that the average fixation duration ranges from 150 to 350 ms [48], if a fixation contains over twice the minimum number of samples required for a fixation (which is set relative to the sampling rate of the webcam), it can be split. This was done to account for potential inaccuracies in the spatial measurements that may cause a fixation too long. While there may be genuine fixations of this length we evaluated these to be the minority. The algorithm selects a split that results in the largest distance between the two new fixations without violating any additional constraints (e.g., each new fixation must still have at least the minimum number of samples). Though there is a risk that this method will inflate the number of fixations detected, we posit that due to the lower sampling rate of video compared to PCCR eye tracking, this will improve overall accuracy.

Calculating Saccades. Once fixations have been approximated, saccades are computed as a straight-line path between sequential fixations using the centroids of each respective fixation as start and end points. For each saccade, we calculate

the distance and angle between start point and end point. Thus, saccade detection relies on accurate fixation detection, which is common in low-frequency eye tracking [8, 20, 58].

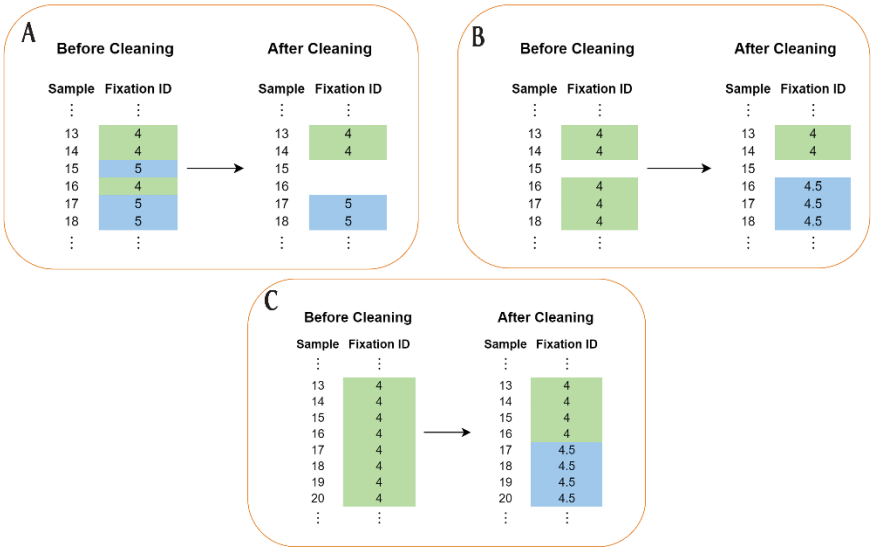

Figure 3. Subfigures A, B, and C show three separate noise removal cases. (A) Example 1 – Overlapping Fixations. (B) Example 2 – Handling noise within a fixation, (C) , Example 3 – Splitting a long fixation.

3 DATASET 1 – PREDICTING MIND WANDERING DURING READING

3.1 Motivation

The first evaluation dataset combines our approach to data collected in the lab as participants completed a reading task [7]. The dataset includes video to which we can apply our methodology (as described above) and gaze data collected with a research-grade PCCR gaze tracker for comparison/evaluation. This presents two advantages for evaluating our fixation estimations. Firstly, by using data collected in a controlled environment, we can be confident that there is minimal noise and thus reliable 'ground truth' in the PCCR data. Second, there has been extensive research examining eye movements in the context of reading [16, 48, 49], meaning that gaze patterns are well understood.

We define accuracy by comparing webcam-based fixation features to those computed with the PCCR gaze tracker, and define useful if the data can be used for a user modeling task, in this case, mind wandering (MW). We focus on MW as it occurs frequently and negatively correlates with cognitive tasks, including learning [14]. Eye gaze from PCCR-based trackers has also been used for MW detection [20, 27], thereby providing a useful comparison.

3.2 Method

Data. This data was previously collected in [7], to assist interpretation we provide summary details here. A total of 152 participants read an educational text and self-reported MW during reading. The text consisted of 57 pages from *Soap Bubbles: Their Colors and the Forces that Mold Them* [10], which was presented on-screen one page (about 114 words see Figure 4) at a time. Videos of participants' faces and upper bodies were recorded with a consumer-grade Logitech C270 webcam at 12.5 frames per second (due to processor constraints of collecting this datastream alongside eye gaze),

whereas participant's eye gaze was simultaneously recorded by the Tobii TX300, a research-grade PCCR gaze tracker recording at 120Hz. At the time of data recording, it was not anticipated that we would extract eye gaze, thus 12.5Hz was considered satisfactory. Participants reported when they caught themselves MW via a key-press, which is a common and validated method to measure an internal conscious state like MW [21, 47]. Pages containing one or more MW reports were considered positive instances of MW. In contrast, pages without a report were negative MW instances.

Computing Gaze Features. Fourteen participants were removed due to video recording errors, leaving 138 participants. We used the method discussed in Section 2 to estimate fixations and saccades from the webcam data. Hyperparameters were tuned using a grid search with search values selected relative to the sampling rate and literature. For example, the minimum number of samples parameter was selected relative to sample rate to ensure that no fixation would be too short relative to what is already known about eye movements. Final hyperparameter values are shown in Table 1. Results were stable across folds, This was somewhat expected, as the hyperparameters are closely linked to the data collection environment. Within each dataset, the environment was static (e.g., same lab). Therefore, the hyperparameters should also be static within the dataset. Fixations and saccades were also estimated from the PCCR eye gaze data using a dispersion-based fixation filter from the Open Gaze and Mouse Analyser algorithm (OGAMA) [58]. In this algorithm, fixations are defined as consecutive samples within a range of 57 pixels (approximately 1 degree of visual angle); saccades are computed from the fixations similar to our method.

We segmented the learning session by each page of reading. Of the potential 7866 pages, 1,867 had insufficient data from either the webcam (449 pages) or PCCR tracker (1,418 pages), leaving 5,999 pages for analysis. This data loss is within the range of typical eye tracking studies [7, 27]. For the webcam gaze estimations, we calculated the following eight descriptive statistics of fixation duration: number, mean, median, minimum, maximum, standard deviation, kurtosis, and skew. We also include the proportion of horizontal saccades, defined as the proportion of saccades with relative angles ≤ 30 degrees above or below the horizontal axis; this measure was included since it approximates standard left-right reading patterns [54]. For this initial proof of concept, we focused on fixation features. We were aware that the fixation centroid estimates were likely to be inaccurate and that this could have a follow-on effect to saccades.

For PCCR gaze data, we compute fifty-seven features used in previous studies on MW detection [7, 20, 27]. These included six general measures based on fixations and saccades (Fixation Duration, Saccade Duration, Saccade Length, Absolute Saccade Angle, Saccade Angle relative to previous Saccade, Saccade Velocity). For these gaze measures, we calculated the same eight descriptive statistics from the distributions of each measure across the window, yielding 54 features. We also included three other features: Fixation Dispersion, Horizontal Saccade Ratio, and Fixation saccade ratio (more detail in [27]). We split these into two feature sets for evaluation purposes, the first (referred to below as PCCR) with only the nine equivalent features contained in the webcam gaze dataset, and the second (referred to below as PCCR Complete) with the full set of 57 features.

Predicting MW. We explored a variety of supervised learning techniques using the sci-kit learn library [45], including logistic regression, decision trees, gradient boosted classifier, support vector machines, and random forest. Participant self-reports of MW served as the ground-truth labels for supervised learning using all three feature sets. Logistic regression proved most successful in all cases; only results for these models are reported here.

Models were trained using user-independent nested 4-fold cross-validation. This meant that data was split so that no participant's instances could be simultaneously in both the training and testing sets. This was done to reduce overfitting and thus improve generalizability to new users. Hyperparameters (when appropriate) were tuned using a grid search nested within the training fold, again using a user-independent scheme. Participants reported mind wandering for 32%

of the 5999 instances, resulting in data skew. To compensate for this class imbalance, we applied the SMOTE algorithm [9] to the training set *only*, without altering the testing set.

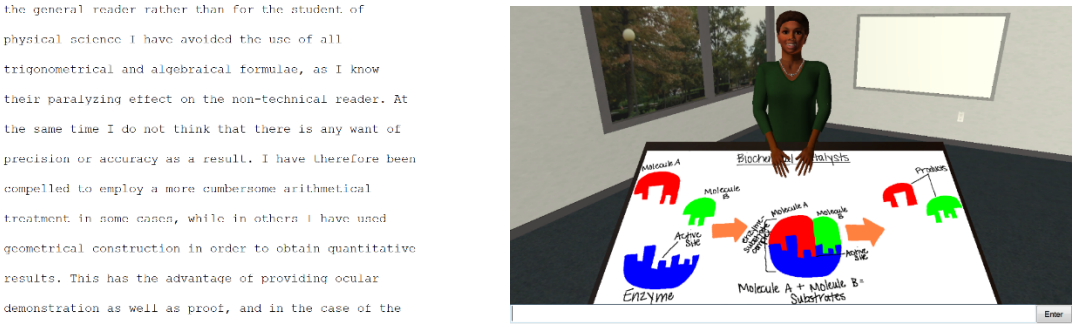

Figure 4. Stimuli Examples. (Left) Screenshot of one page of text as presented for Studies 1 and 2. (Right) Screenshot of the GuruTutor environment used in to collect Dataset 3

3.3 Results

Table 1 provides a summary of the results of all three studies.

Comparing Web-cam and PCCR Gaze Features. We evaluated accuracy of the features estimated from the webcam to those from the PCCR tracker using Spearman correlations, which focuses on rank ordering and is a nonparametric measure. We selected correlations rather than an absolute measure because we are primarily interested in the shared variance between the two methods to compute each feature rather than the absolute difference in feature values. Using Cohen's criterion of .1, .3, and .5 to denote low, medium, or high correlations [12], we found that five features had moderate to strong correlations (rhos between .41 to .50), two were small (rhos of .16), and two had null correlations (rhos of -.03 and -.11).

Predicting Mind Wandering. We evaluate the models using Area Under the Receiver Operator Curve (AUROC). An AUROC of 0.5 represents chance classification, while an AUROC of 1 represents perfect classification. ROC curves are shown in Figure 5. We used DeLongs test for two ROC curves to statistically compare the models as implemented in the pROC package in R [51]. Results indicated above-chance performance for both webcam- and PCCR-based gaze features when using the same set of nine features (AUROC of .59 and .57 respectively, $p = .03$). However, their accuracy was significantly lower ($ps < .001$) than the model trained on the full set of 57 PCCR features (AUROC = .69).

3.4 Discussion

Webcam data faces several challenges for extracting meaningful gaze. We hypothesize that the crucial difference between the PCCR trackers and webcam data collection is one of sampling frequency - the PCCR data was collected at 120Hz, whereas the camera data was collected at approximately 12.5Hz, an order of magnitude difference. Nevertheless, five out of nine webcam gaze features estimations were moderate to strongly correlated with 'ground truth' values calculated from PCCR data ($0.40 < rhos < 0.61$). We also found that, though admittedly modest, a small number of nine features extracted from the webcam could predict MW with above-chance accuracies in a user-independent setting. Importantly, the observed AUROC of .58 is similar to what is achieved by humans (AUROC .60) on the same task [9],

though both were outperformed using the full set of PCCR features. Another important finding was that data loss was lower for webcam data (5.7%) compared to PCCR data (18%), suggesting that video may be more robust (though less accurate) than traditional gaze tracking. Overall, this initial evaluation provides some evidence for the potential for uncalibrated webcam-based eye tracking for user modeling.

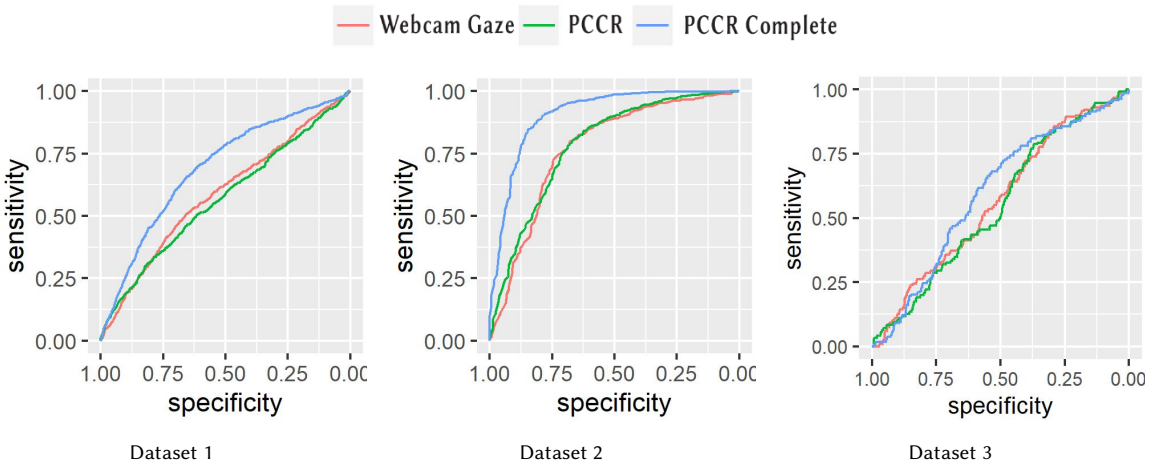

Figure 5. ROC Curves for models trained in Studies 1-3

Table 1. Gaze Estimation Evaluations

	Dataset 1	Dataset 2	Dataset 3
Task Details			
Task	Reading	Reading	Guru
Data Collection Environment	Laboratory	Laboratory	Classroom
PCCR Tracker Used	Tobii TX300	Tobii TX300	EyeTribe
PCCR Sampling Frequency	120Hz	300Hz	30Hz
Participants	138	104	135
Instances	5,999	1,470	1,788
Fixation Approximation Hyperparameters			
Minimum Samples	2	2	2
Spatial Threshold (as measured in the raw gaze vector)	0.02	0.02	0.03
Temporal Threshold	240ms	240ms	320ms
Feature Correlations Between Webcam Derived and PCCR			
Number of Fixations	0.60	0.75	0.23
Mean Fixation Duration	0.49	0.53	0.17
Median Fixation Duration	0.41	0.49	0.21
Fixation Duration Standard Deviation	0.50	0.50	0.23
Minimum Fixation Duration	-0.11	0.12	0.10
Maximum Fixation Duration	0.51	0.52	0.21
Fixation Duration Skew	0.16	0.23	0.10

Fixation Duration Kurtosis	0.16	0.23	0.08
Horizontal Saccade Proportion	-0.03	-0.11	0.06
Modeling			
Construct Modeled	MW	Comprehension	MW
Report Method	Self Caught	Question	Probed
Webcam Gaze Estimations AUROC	0.59	0.77	0.56
PCCR AUROC	0.57	0.79	0.55
PCCR Complete AUROC	0.69	0.91	0.60

Note. Significance is not reported for correlations as instances are not independent

4 DATASET 2 - PREDICTING READING COMPREHENSION

4.1 Motivation

Our analysis of Dataset 2 aimed to serve as a conceptual replication to Dataset 1 in that it also focused on a reading task conducted in the lab on the same equipment. However, ground truth labels in Dataset 1 used for user modeling were for mind wandering, which is highly internal and relies on users being aware of their own attentional state for accurate reporting. Dataset 2 contains modeling labels for participant comprehension, which can be measured through more objective knowledge assessments (i.e., a quiz). Eye gaze has been previously used for assessing comprehension outcomes [13, 54]. Of particular note is prior work from D'Mello et al. [16] that used eye gaze features and reading time to predict reading comprehension outcomes with high accuracy (AUROC scores of 0.9). We use the same data to examine the potential of webcam-based eye gaze estimates at this task.

4.2 Method

This Dataset was originally collected in [15] but is described here for ease of interpretation. Participants were 104 students at a private Midwestern university in the U.S. who participated in exchange for research participation credit. Participants were unlikely to have enrolled in any courses being run by the researchers and had a number of options of how to earn the credit (i.e., students were not required to take our study).The text used was the same as in Dataset 1, with a similar delivery mechanism. Comprehension was assessed during reading using four-option multiple-choice questions that tapped page-specific, textbase-level (i.e., factual) content of the text. Questions could occur after reading any of the 57 pages (apart from the first two), but the number of questions (M = 15, SD = 4) and exact pages on which questions appeared differed by participant.

The original experiment had two experimental groups that defined when questions were delivered [15]. These groups were merged in our analysis as both received comprehension questions. Additionally, if an incorrect answer was provided, participants could re-read the previous page and answer a second question. For this work we consider only the first question asked on each page. Participants' eye gaze was recorded using the Tobii TX300, sampling at 300Hz. Videos of the participant's faces were recorded using a Logitech C270 (again placed above the monitor) at approximately 12.5 frames per second. As in earlier studies, participants were unconstrained and free to move their heads.

There were a total of 1,618 pages with accompanying questions. We removed instances (pages) with missing data from either modality (3.1% missing for PCCR tracker, 3.7% missing for webcam), resulting in 1,470 instances used for modeling. Feature sets were identical to those used in analyzing Dataset 1, as were machine learning procedures, except that the participant's first response accuracy (correct [69%] or incorrect [31%]) to multiple-choice questions served as the ground truth label with features computed from the corresponding page (i.e., the page they just read).

4.3 Results

Comparing Web-cam and PCCR Gaze Features. In general, the correlations between webcam-derived features and PCCR-derived features were somewhat higher than Dataset 1 (see Table 1). Specifically, five yielded strong effects (rhos between .49 and .75), three were small to medium effects (rhos between .12 and .23), and only one resulted in a null effect (rho of -.11).

Predicting Comprehension. The results, shown in Table 1, indicate that all models perform better than chance, with webcam gaze estimations (AUROC = .77) achieving statistically equivalent (p = .23) performance to PCCR features (AUROC = .79), with both being lower than PCCR Complete features (AUROC = .91; ps < .001). Figure 5 shows the ROC curves for each of the three models.

4.4 Discussion

Evaluation of Dataset 2 replicated findings from Dataset 1 in that webcam-based gaze features correlated moderately to highlight with PCCR-based features. However, the accuracy of user modeling (in this case, predicting comprehension) was considerably better than Dataset 1 (which attempted to predict MW), presumably because the ground truth measure was more objective and thereby more reliable than self-reported MW. Thus, Dataset 2 supports the claim that fixation estimations from uncalibrated webcam data are reliable and may be an alternative or complement for a PCCR eye tracker depending on the application domain. In contrast to Dataset 1, data loss in this dataset was approximately equivalent (~3%) for the two modalities.

5 DATASET 3 – PREDICTING MW DURING LEARNING WITH AN INTELLIGENT TUTOR IN THE WILD

5.1 Motivation

The evaluations discussed using Datasets 1 and 2 show promise for uncalibrated webcam-based eye gaze estimations collected in the lab, but how do they perform on data collected in-the wild, with a more interactive task? Dataset 3 was collected in a high school classroom, as students interacted with an Intelligent Tutoring System (ITS) called GuruTutor [40] (see Figure 4). GuruTutor provides a more complex stimulus that supports multimedia output (media and voice) and additional student interactivity compared to reading. As in Dataset 1, this dataset contained user modeling labels for MW. Another key difference with this dataset is that data was collected from an entire classroom of students simultaneously using COTS-eye tracking, thereby providing a much less reliable measure of ground truth eye gaze compared to the lab-based, research-grade PCCR data in earlier Datasets.

5.2 Method

Dataset 3 is the same as that used by Hutt et al. in [27], which provides full details. Participants were 135 high school students in their regular biology class. GuruTutor (Guru) is an ITS designed to teach biology topics through collaborative conversations in natural language. It was modeled after interactions with expert human tutors and has been shown to be effective at promoting learning and retention at levels similar to group tutoring by human tutors. Guru engages the student through natural language conversations, using an animated tutor agent that references a multimedia workspace. Students engage in multiple activities throughout the learning session, including common ground building (a collaborative lecture), dialogue sessions (where the tutor asks students questions), and completing a concept map. Participants each completed two Guru sessions with MW measured with thought probes occurring every 90-120 seconds [36]. Specifically, the probes consisted of an auditory beep along with an opaque overlay on screen that paused the

tutoring session to solicit MW responses from participants. Participants received on average 12 probes over the course of each session (2,720 probes in all) with a mean MW rate of 28% (SD = 24%, min = 0%, max = 100%).

Throughout the session, eye movements were recorded using the EyeTribe, a COTS PCCR tracker, retailing for approximately 100-150 USD and sampling at 30Hz. In order to protect student privacy, no video was recorded. Instead, facial, pose, and gaze features were extracted in real-time with a consumer-grade Logitech C270 webcam at 12.5 frames per second. No video was recorded in this study, only the features extracted. The webcam was placed at the top of the monitor, approximately in the center.

We derived features the same sets of three gaze features over a 30 second-window before each MW probe, similar to [27]. To ensure a fair comparison, we only consider cases with valid data from both the webcam and the PCCR Tracker. A total of 454 (16.6%) instances had missing data for the webcam, whereas 547 instances (20.1%) were missing for PCCR tracking

5.3 Results

Comparing Web-cam and PCCR Gaze Features. We found that correlations were lower than in Datasets 1 and 2 (see Table 1), ranging from 0.06 to 0.23. Specifically, five resulted in small to medium effects (rhos between .17 and .23), two small effects (rhos of .1), and three null effects (rhos of .06 and .08). This is potentially due to the lower sampling rate increased noise for the comparison PCCR tracker.

Predicting MW. Models were again evaluated using AUROC and DeLongs test for comparisons. The results (see Figure 5 and Table 1) indicate that all models performed better than chance, with webcam gaze estimations (AUROC= .56) providing a nonsignificant ($p = .50$) difference in accuracy over PCCR features (AUROC= .55). Webcam gaze features were lower (though not significantly, $p = .18$) than PCCR Complete features (AUROC = .60).

5.4 Discussion

We have shown that the results from earlier Datasets generalize to this authentic, real-world environment. In this case, the correlations between webcam- and PCCR-computed features were lower than in Datasets 1 and 2, potentially due to inaccuracy in both data streams as the COTS tracker is also known to provide noisy, inaccurate data. Additionally, the environment contained more distractions and was less controlled than a laboratory (i.e., Datasets 1 and 2).

We have shown that webcam gaze estimations can be used for MW detection in-the-wild above chance and are comparable (or even slightly better) than the equivalent PCCR data. The slight improvement observed in the webcam model is surprising. One potential explanation is that the video data is slightly more robust to head movements (as long as the face is still visible) than the COTS tracker, resulting in more robust data. This would also support why less data loss was observed in both this dataset and Dataset 1 for webcam vs. the PCCR tracker. However, given that there was no significant difference in the overall accuracy of the models, it is unlikely this is a substantial difference. As in earlier analysis, data from the full 57 gaze features achieved a higher AUROC score than the limited nine features. This finding is in accordance with the theory that additional feature engineering (if the data is sufficiently accurate) could improve user modeling outcomes.

6 GENERAL DISCUSSION

This paper developed an unsupervised algorithm that maps uncalibrated three-dimensional gaze vectors from a webcam/video data to higher-order eye gaze fixation features used in user modeling applications, bypassing the need for a traditional PCCR eye tracker and a calibration process. We evaluated our methodology over three datasets with

data collected in the laboratory with research-grade eye tracking and in-the-wild with COTS eye tracking. Our main findings are summarized below, followed by a discussion of applications and limitations.

6.1 Main Findings

Our two performance criteria included: (1) correlation of webcam-based features to features computed from a traditional gaze tracker; (2) performance in user modeling tasks compared to a PCCR eye tracker using the same set of features available from the webcam and an expanded set from the literature. By using three diverse datasets, we were able to show initial feasibility of this approach. Specifically, estimated gaze features were generally positively correlated with corresponding values calculated from PCCR data in most cases. Five out of the nine features (number of fixations, mean, median, standard deviation, and maximum fixation duration) generally resulted in moderate to large correlations, two others (skew and kurtosis of fixation duration) yielded small correlations, whereas minimum fixation duration and horizontal saccade proportion resulted in null correlations. Thus, the method is more accurate for estimations measures of the central tendency of the fixation durations. It is also very inaccurate at measuring saccadic features, as evidenced by the low accuracy of estimating the proportion of horizontal saccades. Further, as expected, correlations were higher for the laboratory studies with research-grade PCCR eye tracking as a ground truth measure (Datasets 1 and 2) compared to a real-world study with a PCCR eye tracker for comparison (Dataset 3). Nevertheless, several of the gaze features could still be computed with nonzero accuracy, even in Dataset 3. Thus, although the gaze estimations from traditional RGB cameras are known to have a high error when translated to screen coordinates [5, 33, 43], the proposed method bypasses this step by directly mapping to the feature space.

Our results also indicate that webcam-based gaze features could be used in downstream user modeling tasks, predicting mind wandering and reading comprehension. Specifically, we found that gaze fixation approximations from video data could predict comprehension with moderate accuracy (AUROC of .77) and significantly above chance. MW is a more complex construct for modeling purposes due to the reliance on self-reports for ground truth values. Despite this, we show that we can detect MW during both reading and interactions with an intelligent tutoring system. The accuracies achieved in these cases (AUROCs of .58 and .56), though notably lower, are still above chance and comparable to those achieved from the same feature obtained with a PCCR eye tracker, which highlights the complexity of the task from a very limited feature set. To this point, a recent study found that individual human judges achieved an AUROC of .55 on a task of judging MW from short 10-sec video clips of users engaged in a computerized reading task [9].

That said, when compared to the full set of 57 PCCR features, accuracies obtained from the limited set of 9 features were lower in all cases. One reason for this might be due to the exclusion of features related to saccades, which comprised a majority of the remaining 57 features. Saccades are presumably as important as gaze fixations, but were currently not accurately computable from webcam data alone, suggesting that webcam gaze might not yet be able to completely replace a PCCR tracker when overall accuracies are low (mind wandering) but might be an alternative when accuracies are generally high (reading comprehension).

This work is an initial exploration of new opportunities for gaze tracking. While some of the results presented may be context-specific, this work shows both feasibility and promise for the approach presented. Future work (discussed more below) will allow for further development of this methodology and more opportunities for scalable gaze tracking.

6.2 Applications

The principal advantage of this work is to leverage eye tracking (and the decades of research that have utilized it) for multimodal interaction/analysis at scale and in environments where calibration may be undesirable or impossible (e.g.,

limited time interactions, public spaces, post-hoc analysis). Of course, the utility of the approach will depend on the desired accuracy of the application. As expected, we see inaccuracies in estimations derived from webcam data however it is encouraging to note that despite this, the data is still suitable for some user modeling tasks. All applications must consider the potential that inaccuracy might have on measurements. For example, if using models for intervention, said interventions should be fail-soft (i.e., they are not harmful if delivered based on incorrect sensing).

Webcam-based gaze estimation might also be useful when robustness to missing data is key. One unexpected finding was that data loss from the webcam was much lower than the PCCR eye tracker for Dataset 1, somewhat lower for Dataset 2, and about the same for Dataset 3. Thus, though less accurate and expansive than PCCR-based tracking, webcam-based estimation might provide a more robust method. This suggests that there might be merits to combining the two (e.g., using PCCR when available and webcam when not) when robustness is important.

This work is also of potential interest to researchers. Our analysis has shown, firstly, that useful gaze data can be extracted in a posthoc manner, even if there was no formal calibration. This approach may thus present opportunities to add an additional modality of analysis to previously collected data. Similarly, we have shown that the webcam can provide a cost-effective method for collecting gaze data without calibration for future data collections.

In all potential applications, it is critical to address privacy. Specifically, users must be informed of what data is collected and how it will be used. The informed consent and assent procedure used in research may not be possible in real-world environments. However, it is important to note that this method does not require video to be recorded. Instead, features can be extracted in real-time (as was done in Dataset 3) and the video discarded, which reduces privacy concerns and legal issues [34, 39, 44] (though not negating the need for transparent communication with users).

6.3 Limitations and Future Work

This work is not without limitations. Firstly, we did not consider individual differences, most notably whether a user wears glasses. Early work using RGB cameras for gaze estimations showed that glasses could sometimes reduce accuracy due to additional reflections [28], which have not been separately examined here. A further critical individual difference to examine is race/ethnicity. Previous work has shown that facial recognition algorithms can be biased towards certain races [22], impacting detection accuracy. OpenFace was developed and evaluated using the Labeled Faces in the Wild dataset [32], which exhibits "natural" variability in age, gender, race, and other factors. Nevertheless, future work must consider if the performance of our approach is in any way impacted by race.

Another limitation is that it is not clear if the source of errors is the proposed methodology or OpenFace detection of eye gaze. Future work should investigate this further to understand how error can be reduced. One option would be to compare uncalibrated webcam data to calibrated webcam data, such as in [64]. There have also been recent successes with deep-learning appearance-based gaze estimations [17, 63]. Combined with the clustering approach here, these could yield more accurate results.

The fact that the clustering approach requires hyperparameter tuning is also a potential limitation, though we note that the minimum number of samples and spatial threshold hyperparameters were quite similar across the three studies (Table 1), suggesting they might be internal parameters. The temporal threshold differed for Dataset 3, however, this parameter can be adjusted based on the sampling rate of the comparison PCCR tracker, which is a known value. Thus, future work should examine whether these parameters can be preselected and subsequently automatically adjusted.

The low sampling rate of a traditional webcam (often capped at 30Hz) also presents a limitation. Detailed eye movements such as saccades and smooth pursuits (following a moving object) require a much higher sampling frequency, meaning it is unlikely they can ever be accurately detected through this approach. However, many COTS

PCCR trackers also encounter this issue, and lower level features such as the number of fixations have been consistently shown to be a rich source of input. All video data considered here was recorded at approximately 12.5Hz due to processor constraints; future work will examine if increasing to 30Hz (or 60Hz where possible) increases estimation accuracy. Alongside this work, we can also explore the models further to examine similarity in how features are used to generate predictions and consider similarities and differences in the learnt models as a further evaluation of accuracy.

Finally, the measure of comprehension used in Dataset 2 considered a very shallow model of comprehension, with questions asked immediately after a page of reading. Research has shown that these kinds of questions cannot assess long-term understanding, but are instead assessing short-term comprehension [37]. Further investigation is required to assess how this approach might be used to address long-term (or deep) comprehension.

6.4 Concluding Remarks

We developed and evaluated an unsupervised algorithm to approximate gaze fixation features from raw gaze vectors derived from webcams/video data. Our results show that the computed features are moderately accurate when compared to traditional PCCR eye trackers and can be used to model complex constructs such as mind wandering and reading comprehension, both in the lab and in the wild. They also highlighted future research opportunities to scale gaze-based interactions to new environments and larger populations without requiring specialized equipment and/or tedious calibration procedures.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation (NSF) (DRL 1920510 and IIS 1523091). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF. Thanks to fellow lab members for their assistance in the data collection.

REFERENCES

[1] Alnajar, F., Gevers, T., Valenti, R. and Ghebreab, S. 2013. Calibration-free gaze estimation using human gaze patterns. *Proceedings of the IEEE international conference on computer vision* (2013), 137–144.

[2] Bacivarov, I., Ionita, M. and Corcoran, P. 2008. Statistical Models of Appearance for Eye Tracking and Eye-Blink Detection and Measurement. *IEEE Transactions on Consumer Electronics*. 54, 3 (2008), 1312–1328.https://doi.org/10.1109/TCE.2008.4637622.

[3] Bal, E., Harden, E., Lamb, D., Van Hecke, A.V., Denver, J.W. and Porges, S.W. 2010. Emotion Recognition in Children with Autism Spectrum Disorders: Relations to Eye Gaze and Autonomic State. *Journal of Autism and Developmental Disorders*. 40, 3 (2010), 358–370.

[4] Baltrušaitis, T., Robinson, P. and Morency, L.P. 2013. Constrained local neural fields for robust facial landmark detection in the wild. *Proceedings of the IEEE International Conference on Computer Vision* (2013).https://doi.org/10.1109/ICCVW.2013.54.

[5] Baltrusaitis, T., Zadeh, A., Lim, Y.C. and Morency, L.P. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*. (May 2018), 59–66.https://doi.org/10.1109/FG.2018.00019.

[6] Baluja, S. and Pomerleau, D. 1994. Non-Intrusive Gaze Tracking Using Artificial Neural Networks, CMU-CS-94-102. *Neural Networks*. (1994), 753–760.

[7] Bixler, R. and D'Mello, S.K. 2016. Automatic Gaze-Based User-Independent Detection of Mind Wandering during Computerized Reading. *User Modeling and User-Adapted Interaction*. 26, 1 (2016), 33–68.https://doi.org/10.1007/s11257-015-9167-1.

[8] Bixler, R. and D'Mello, S.K. 2014. Toward Fully Automated Person-Independent Detection of Mind Wandering. *User Modeling Adaptation and Personalization* (Aalborg, Denmark, Denmark, 2014), 37–48.https://doi.org/10.1007/978-3-319-08786-3_4.

[9] Bosch, N. and D'Mello, S.K. 2022. Can Computers Outperform Humans in Detecting User Zone-Outs? Implications for Intelligent Interfaces. *ACM Trans. Comput.-Hum. Interact.* 29, 2 (Jan. 2022).https://doi.org/10.1145/3481889.

[10] Boys, C. V 1959. Soap Bubbles, Their Colours and the Forces Which Mold Them. (1959).

[11] Capozzi, F. and Ristic, J. 2021. Attentional Gaze Dynamics in Group Interactions. *Visual Cognition*. (2021), 1–16.

[12] Cohen, P., West, S.G. and Aiken, L.S. 2014. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. (2014).

[13] Copeland, L., Gedeon, T. and Mendis, B.S.U. 2014. Predicting Reading Comprehension Scores from Eye Movements Using Artificial Neural Networks and Fuzzy Output Error. *Artif. Intell. Res.* 3, 3 (2014), 35–48.

[14] D'Mello, S.K. 2018. What Do We Think About When We Learn? *Deep Comprehension*. (2018), 52–67.

[15] D'Mello, S.K., Mills, C., Bixler, R. and Bosch, N. 2017. Zone out No More: Mitigating Mind Wandering during Computerized Reading. *International Educational Data Mining Society*.

[16] D'Mello, S.K., Southwell, R. and Gregg, J. 2020. Machine-Learned Computational Models Can Enhance the Study of Text and Discourse: A

Case Study Using Eye Tracking to Model Reading Comprehension. *Discourse Processes*. (2020).https://doi.org/10.1080/0163853X.2020.1739600.
[17] D, M.L.R. and Biswas, P. 2021. Appearance-Based Gaze Estimation Using Attention and Difference Mechanism. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (Jun. 2021), 3143–3152.
[18] Deubel, H. and Schneider, W.X. 1996. Saccade Target Selection and Object Recognition: Evidence for a Common Attentional Mechanism. *Vision Research*. 36, 12 (1996), 1827–1837.https://doi.org/10.1016/0042-6989(95)00294-4.
[19] Ester, Kriegel, Sander and Xu 1996. DBSCAN. *KDD-96 Proceedings*. (1996).https://doi.org/10.1.1.71.1980.
[20] Faber, M., Bixler, R. and D'Mello, S.K. 2018. An Automated Behavioral Measure of Mind Wandering during Computerized Reading. *Behavior Research Methods*. 50, 1 (2018), 134–150.https://doi.org/10.3758/s13428-017-0857-y.
[21] Franklin, M.S., Smallwood, J. and Schooler, J.W. 2011. Catching the Mind in Flight: Using Behavioral Indices to Detect Mindless Reading in Real Time. *Psychonomic Bulletin & Review*. 18, 5 (Oct. 2011), 992–997.https://doi.org/10.3758/s13423-011-0109-6.
[22] Garvie, C. and Frankle, J. 2016. Facial-Recognition Software Might Have a Racial Bias Problem. *The Atlantic*. (2016).
[23] Guestrin, E.D. and Eizenman, M. 2006. General Theory of Remote Gaze Estimation Using the Pupil Center and Corneal Reflections. *IEEE Transactions on Biomedical Engineering*. (2006).https://doi.org/10.1109/TBME.2005.863952.
[24] Hayward, D.A., Voorhies, W., Morris, J.L., Capozzi, F. and Ristic, J. 2017. Staring Reality in the Face: A Comparison of Social Attention Across Laboratory and Real World Measures Suggests Little Common Ground. *Canadian Journal of Experimental Psychology*. (2017).https://doi.org/10.1037/cep0000117.
[25] Hoffman, J.E. and Subramaniam, B. 1995. The Role of Visual Attention in Saccadic Eye Movements. *Perception & Psychophysics*. 57, 6 (1995), 787–795.https://doi.org/10.3758/BF03206794.
[26] Hutt, S., Krasich, K., Brockmole, J.R. and D'Mello, S.K. 2021. Breaking Out of the Lab: Mitigating Ming Wandering with Gaze-Based Attention-Aware Technology in Classrooms. *ACM SIGCHI: Computer-Human Interaction* (2021).
[27] Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., Brockmole, J.R. and D'Mello, S.K. 2019. Automated Gaze-Based Mind Wandering Detection during Computerized Learning in Classrooms. *User Modeling and User-Adapted Interaction*. 29, 4 (Sep. 2019), 821–867.https://doi.org/10.1007/s11257-019-09228-5.
[28] Ince, I.F. and Kim, J.W. 2011. A 2D Eye Gaze Estimation System with Low-Resolution Webcam Images. *Eurasip Journal on Advances in Signal Processing*. 2011, (2011), 1–11.https://doi.org/10.1186/1687-6180-2011-40.
[29] Itier, R.J. and Batty, M. 2009. Neural Bases of Eye and Gaze Processing: The Core of Social Cognition. *Neuroscience \& Biobehavioral Reviews*. 33, 6 (2009), 843–863.
[30] Just, M.A. and Carpenter, P. 1976. Eye Fixations and Cognitive Processes. *Cognitive Psychology*. 8, 4 (1976), 441–480.https://doi.org/10.1016/0010-0285(76)90015-3.
[31] Kar, A. and Corcoran, P. 2017. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*.
[32] Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H. and Hua, G. 2016. Labeled Faces in the Wild: A Survey. *Advances in Face Detection and Facial Image Analysis*. (2016).https://doi.org/10.1007/978-3-319-25958-1_8.
[33] Lin, Y.T., Lin, R.Y., Lin, Y.C. and Lee, G.C. 2013. Real-Time Eye-Gaze Estimation Using a Low-Resolution Webcam. *Multimedia Tools and Applications*. 65, 3 (2013), 543–568.https://doi.org/10.1007/s11042-012-1202-1.
[34] Malhotra, N.K., Kim, S.S. and Agarwal, J. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*.
[35] Mathews, A., Fox, E., Yiend, J. and Calder, A. 2003. The Face of Fear: Effects of Eye Gaze and Emotion on Visual Attention. *Visual Cognition*. 10, 7 (2003), 823–835.
[36] Mills, C., D'Mello, S.K., Bosch, N. and Olney, A.M. 2015. Mind Wandering During Learning with an Intelligent Tutoring System. *Artificial Intelligence in Education* (Madrid, Spain, Spain, Jun. 2015), 267–276.https://doi.org/10.1007/978-3-319-19773-9_27.
[37] Mills, C., Gregg, J.M., Bixler, R. and D'Mello, S.K. 2021. Eye-Mind Reader: An Intelligent Reading Interface That Promotes Long-Term Comprehension by Detecting and Responding to Mind Wandering. *Hum. Comput. Interact.* 36, 4 (2021), 306–332.https://doi.org/10.1080/07370024.2020.1716762.
[38] Müller, P., Huang, M.X., Zhang, X. and Bulling, A. 2018. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. *Eye Tracking Research and Applications Symposium (ETRA)* (2018).https://doi.org/10.1145/3204493.3204549.
[39] Nguyen, D.H., Bedford, A., Bretana, A.G. and Hayes, G.R. 2011. Situating the Concern for Information Privacy through an Empirical Study of Responses to Video Recording. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (2011), 3207–3216.https://doi.org/10.1145/1978942.1979419.
[40] Olney, A.M., D'Mello, S.K., Person, N., Cade, W., Hays, P., Williams, C., Lehman, B. and Graesser, A. 2012. Guru: A Computer Tutor That Models Expert Human Tutors. *Intelligent Tutoring Systems* (Chania, Crete, Greece, Jun. 2012), 256–261.https://doi.org/10.1007/978-3-642-30950-2_32.
[41] Olney, A.M., Person, N.K. and Graesser, A.C. 2012. Guru: Designing a Conversational Expert Intelligent Tutoring System. *Cross-Disciplinary Advances in Applied Natural Language Processing*. (2012), 156–171.https://doi.org/10.4018/978-1-61350-447-5.ch011.
[42] Papoutsaki, A., Laskey, J. and Huang, J. 2017. SearchGazer: Webcam Eye Tracking for Remote Studies of Web Search. *Proceedings of the 2017 Conference Human Information Interaction and Retrieval* (2017), 17–26.https://doi.org/10.1145/3020165.3020170.
[43] Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J. and Hays, J. 2016. WebGazer : Scalable Webcam Eye Tracking Using User Interactions. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (2016), 3839–3845.
[44] Paul, C., Scheibe, K. and Nilakanta, S. 2020. Privacy Concerns Regarding Wearable IoT Devices: How it is Influenced by GDPR? *Proceedings of the 53rd Hawaii International Conference on System Sciences* (2020), 4388–4397.https://doi.org/10.24251/hicss.2020.536.
[45] Pedregosa, F. et al. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12, (2011), 2825–2830.https://doi.org/10.1007/s13398-014-0173-7.2.
[46] Ramadan, S., Abd-almageed, W. and Smith, C.E. 2002. Eye Tracking Using Active Deformable Models. December (2002).
[47] Randall, J.G., Oswald, F.L. and Beier, M.E. 2014. Mind-Wandering, Cognition, and Performance: A Theory-Driven Meta-Analysis of Attention Regulation. *Psychological Bulletin*. 140, 6 (Nov. 2014), 1411–1431.https://doi.org/10.1037/a0037428.
[48] Rayner, K. 2009. Eye Movements and Attention in Reading, Scene Perception, and Visual Search. *Quarterly Journal of Experimental Psychology*. 62, 8 (2009).https://doi.org/10.1080/17470210902816461.
[49] Rayner, K. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*. 124, 3 (Nov. 1998), 372–422.https://doi.org/10.1037/0033-2909.124.3.372.
[50] Reinders, M.J.T. 2014. Eye Tracking by Template Matching Using an Automatic Codebook Generation Scheme. January 1997 (2014).

[51] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. and Müller, M. 2011. PROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinformatics*. 12, 1 (2011), 1–8.

[52] Schubert, E., Sander, J., Ester, M., Kriegel, H.P. and Xu, X. 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems*. (2017).https://doi.org/10.1145/3068335.

[53] Smallwood, J. and Schooler, J.W. 2006. The Restless Mind. *Psychological Bulletin*. 132, 6 (Nov. 2006), 946–958.https://doi.org/10.1037/0033-2909.132.6.946.

[54] Southwell, R., Gregg, J., Bixler, R. and D'Mello, S.K. 2020. What Eye Movements Reveal About Later Comprehension of Long Connected Texts. *Cognitive Science*. 44, 10 (2020), e12905.

[55] Sugano, Y., Matsushita, Y. and Sato, Y. 2012. Appearance-Based Gaze Estimation Using Visual Saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 35, 2 (2012), 329–341.

[56] Tran, M., Sen, T., Haut, K., Ali, M.R. and Hoque, M.E. 2020. Are You Really Looking at Me? A Feature-Extraction Framework for Estimating Interpersonal Eye Gaze from Conventional Video. *IEEE Transactions on Affective Computing*. (2020), 1.https://doi.org/10.1109/TAFFC.2020.2979440.

[57] Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K. and Navalpakkam, V. 2020. Accelerating Eye Movement Research via Accurate and Affordable Smartphone Eye Tracking. *Nature Communications*. 11, 1 (2020), 4553.https://doi.org/10.1038/s41467-020-18360-5.

[58] Voßkühler, A., Nordmeier, V., Kuchinke, L. and Jacobs, A.M. 2008. OGAMA (Open Gaze and Mouse Analyzer): Open-Source Software Designed to Analyze Eye and Mouse Movements in Slideshow Study Designs. *Behavior Research Methods*. (2008).https://doi.org/10.3758/BRM.40.4.1150.

[59] Wood, E., Baltruaitis, T., Zhang, X., Sugano, Y., Robinson, P. and Bulling, A. 2015. Rendering of eyes for eye-shape registration and gaze estimation. *Proceedings of the IEEE International Conference on Computer Vision* (2015).https://doi.org/10.1109/ICCV.2015.428.

[60] Wu, Y.L., Yeh, C.T., Hung, W.C. and Tang, C.Y. 2014. Gaze Direction Estimation Using Support Vector Machine with Active Appearance Model. *Multimedia Tools and Applications*. 70, 3 (2014), 2037–2062.https://doi.org/10.1007/s11042-012-1220-z.

[61] Yonetani, R., Kawashima, H. and Matsuyama, T. 2012. Multi-mode Saliency Dynamics Model for Analyzing Gaze and Attention. *Proceedings of the Symposium on Eye Tracking Research and Applications* (New York, NY, USA, 2012), 115–122.https://doi.org/10.1145/2168556.2168574.

[62] Zhang, X., Sugano, Y. and Bulling, A. 2017. Everyday eye contact detection using unsupervised gaze target discovery. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (2017).https://doi.org/10.1145/3126594.3126614.

[63] Zhang, X., Sugano, Y., Fritz, M. and Bulling, A. 2019. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 41, 1 (2019), 162–175.https://doi.org/10.1109/TPAMI.2017.2778103.

[64] Zhao, Y., Lofi, C. and Hauff, C. 2017. Scalable Mind-Wandering Detection for MOOCs: A Webcam-Based Approach. *Data Driven Approaches in Digital Education - 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia, September 12-15, 2017, Proceedings* (2017), 330–344.https://doi.org/10.1007/978-3-319-66610-5_24.