Moment-to-Moment Detection of Internal Thought during Video Viewing from Eye Vergence Behavior

Michael Xuelin Huang
Max Planck Institute for Informatics
Saarland Informatics Campus
mhuang@mpi-inf.mpg.de

Jiajia Li
School of Art and Design
Guangdong University of Technology
lijiajia.simg@gmail.com

Grace Ngai, Hong Va Leong
Department of Computing
Hong Kong Polytechnic University
{csgngai,cshleong}@comp.polyu.edu.hk

Andreas Bulling
Institute for Visualisation and Interactive Systems
University of Stuttgart
andreas.bulling@vis.uni-stuttgart.de

ABSTRACT

Internal thought refers to the process of directing attention away from a primary visual task to internal cognitive processing. It is pervasive and closely related to primary task performance. As such, automatic detection of internal thought has significant potential for user modeling in human-computer interaction and multimedia applications. Despite the close link between the eyes and the human mind, only few studies have investigated vergence behavior during internal thought and none has studied moment-to-moment detection of internal thought from gaze. While prior studies relied on long-term data analysis and required a large number of gaze characteristics, we describe a novel method that is user-independent, computationally light-weight and only requires eye vergence information readily available from binocular eye trackers. We further propose a novel paradigm to obtain ground truth internal thought annotations by exploiting human blur perception. We evaluated our method during natural viewing of lecture videos and achieved a 12.1% improvement over the state of the art. These results demonstrate the effectiveness and robustness of vergence-based detection of internal thought and, as such, open new research directions for attention-aware interfaces.

CCS CONCEPTS

Human-centered computing→ Human computer interaction (HCI)

KEYWORDS

Experimental paradigm; mind wandering; attention shift

ACM Reference format:

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00.
DOI: https://doi.org/10.1145/3343031.3350573

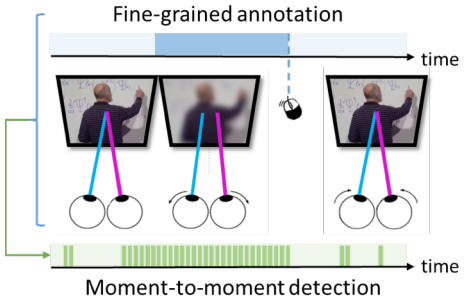


Figure 1. We propose a novel experimental paradigm for fine-grained internal thought annotation and a new method that performs moment-to-moment detection of internal thought from eye vergence behavior.

Michael Xuelin Huang, Jiajia Li, Grace Ngai, Hong Va Leong, and Andreas Bulling. 2019. Moment-to-Moment Detection of Internal Thought during Video Viewing from Eye Vergence Behavior. In *Proceedings of ACM Multimedia conference (MM'19), October 21-25, Nice, France.* ACM. https://doi.org/10.1145/3343031.3350573.

1 Introduction

While a large number of works have studied attention shifts between external stimuli, shifts from external stimuli to internal thought has only been studied rather recently [35]. The *attention shift* in this paper refers to the *shift of attentional focus away from the visual task towards users' internal thoughts*. Specifically, we define the visual task as attending to the visual material and catching every conveyed message. In practice, users' *internal thoughts* can be task-related, e.g. goal-directed thought, as well as task-unrelated, i.e. *mind wandering* [5][31][33]. Although internal thoughts can be positive to understand the visual material, it is also highly important for an intelligent system to immediately be aware of users' attention shift, so as to allow them to review the important missing information or to build the users' attentional profiles. Moreover, mind wandering, a common form of internal thoughts, was showed to be pervasive (over 50% of the time)

during everyday activities [19] and it oftentimes decreases learning performance [25]. As such, *moment-to-moment detection* of internal thought, i.e. fast detection with latency no more than a second, is an essential but currently missing technique towards understanding and improving user experiences, for example via attention (re)direction [7] and task rearrangement [8].

Previous work has explored attention detection based on gaze patterns and pupil dilation. However, the vast majority of them focused on reading [2][11][14][21] for which gaze patterns are rather unique. In contrast, far fewer works have aimed to detect attention in other contexts, such as movie watching [23] or during interactions with a tutoring system [18], and no effort has been made to achieve moment-to-moment detection.

We propose two major improvements over the state of the art. First, inspired by recent psychological studies that observed vergence changes (i.e. both eyes either rotating inwards or outwards) during internally directed cognition [1][41], we propose internal thought detection from eye vergence behavior. This is also supported by the phenomenon of "staring into space" [41] and the fact that people in a visually relaxed state may transit their visual focus to a resting state called *tonic vergence* [39]. Unlike methods that relied on long-term eye movement characteristics, such as the number of fixations [2][18] or erratic fixational patterns [27], vergence analysis allows for moment-to-moment detection from a one-second sliding window. It may therefore explain temporal and causal changes of human performance [25], search intents [29][30], and attention [34][43] in various activities. To our knowledge, this is the first work to leverage eye vergence behavior for moment-to-moment internal thought detection.

Second, this paper proposes a novel experimental paradigm for attention studies by providing fine-grained annotation of internal thought (see Figure 1). Although post-hoc self-reports provide an undoubtedly useful and valid measurement of attentional state [32][34], they are generally associated with a predefined time segment, thus lacking information on the precise start and end. Existing methods also rely on participants' awareness of mind state and perception of time, which can be subjective and error-prone [34]. In contrast, we exploit human perception of a *gradual blur effect* to estimate the start and precisely measure the end of internal thought. This method generates reliable and fine-grained annotations for internal thought and opens new research avenues in human-computer interaction and neuroscience communities.

The contributions of our work are three-fold. First, we propose a novel experimental paradigm for internal thought studies without intrusive intervention. Second, we propose the use of eye vergence behavior for light-weight and moment-to-moment detection of internal thought. Third, we conduct experimental evaluations to validate our experimental paradigm and evaluate our method for internal thought detection during video viewing.

2 Related Work

This study is related with prior works on gaze-based attention detection and the link between eye vergence and attention.

2.1 Gaze-Based Attention Detection

The detection of users' attention or inattention is essential to intelligent interfaces. Conati et al. presented an extensive review on the studies that used gaze data to analyze, model and respond to users' attentional states [6]. In practice, users' inattention can either be overt or covert. Overt attention change denotes users intentionally direct visual attention to or away from the task. It is relatively straightforward to detect given eye tracking data. In contrast, detecting covert inattention, e.g. mind wandering, is more challenging but interesting to us.

Most previous works focused on understanding gaze behavior during mind wandering and only a few explored the inverse, i.e. gaze-based mind wandering detection. Franklin et al. made the first attempt in a constrained word-by-word reading paradigm [13]. After that, there was a line of works investigated mind wandering detection based on the statistics of fixations, saccades, blinks, and pupil dilation. Bixler and D'Mello compared content-dependent and -independent gaze features during natural reading [2]. Faber et al. recently suggested that content-independent gaze features with a 12s window was suitable for reading tasks [11]. Mills et al. extended the study to film watching [23]. Hutt et al. investigated interactions with an interactive tutoring system [18], and later on lecture viewing [17]. They also found that content-independent gaze features performed consistently better and that longer windows (20-30s) were preferable for non-reading tasks.

These studies yielded two key insights. First, small windows ($\leq$10s) may contain insufficient fixation and saccade information for covert inattention detection [2][18] and shorter time window produces a lower accuracy [11][17][18]. This limitation constrains the moment-to-moment detection of internal thought and further motivates the use of eye vergence features. Second, content-dependent features contribute little to the classification accuracy across reading, film, and lecture viewing [2][17][23]. Therefore, in this paper, we also focus on content-independent features.

2.2 Eye Vergence and Attention

Prior works have studied eye vergence and attention. Solé Puig et al. suggested that eye vergence is linked with covert spatial attention, where human visual attention is directed to other visual stimuli in peripheral vision without overt eye movement of orienting [36]. Please note that the focus of their study was visuospatial attention (i.e. the spatial shift of visual focus), rather than the change of cognitive focus in the current study. Lately, their further studies also showed that eye vergence behavior reflecting covert spatial attention can be used to identify children with attention deficit hyperactivity disorder [37].

In contrast to the studies on covert spatial attention, two recent psychological studies explored the link between eye vergence and internal cognition, which well supports our idea of vergence-based internal thought detection. Walcher et al. found that internally directed cognition leads to increased variability in eye vergence in a letter-by-letter reading task [41]. Benedek et al. showed a higher variability and smaller eye vergence during internally directed cognition in anagram and sentence generation tasks [1]. However, unlike this study, their objective was to

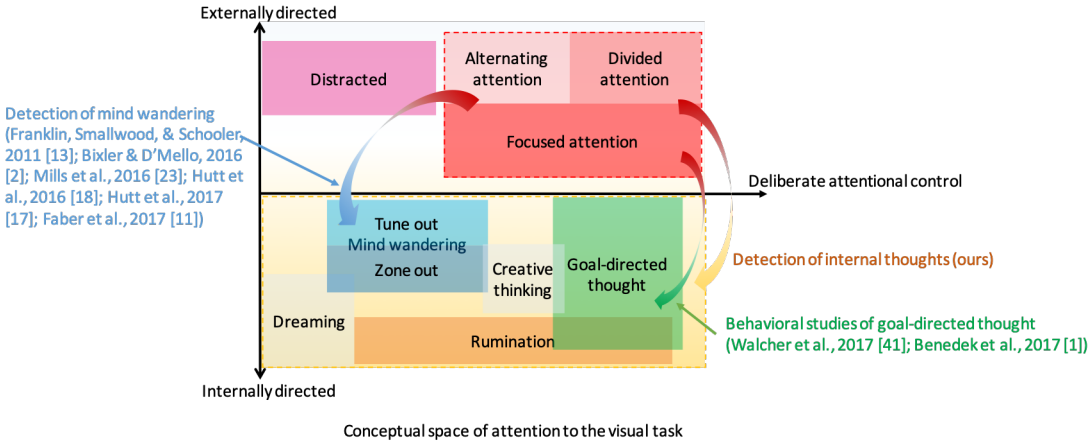


Figure 2. Conceptual space of users' attention to the visual task. We aim to identify attention shift from the visual task (i.e. external stimuli) to internal thought. We also highlight some of the gaze-based studies in the figure.

investigate behavioral patterns during *deliberate* (rather than *spontaneous*) internal cognition, and they did not perform automatic detection. Further, their methods also required extended observation of gaze behavior over 20-30s.

3 Internal Thought Detection in the Conceptual Space of Attention

According to the attentional framework described by D'Mello [7], attention to visual material can take three major forms (see also Figure 2): (1) focused attention: users completely focus on one part of the visual task, e.g. the lecturer in video; (2) alternating attention: users switch attention between different parts of the visual material, e.g. between the lecturer and the notes on board; and (3) divided attention: users attend to visual information and meanwhile process the narration. When attention shifts away from the visual task, it can be either *overtly* distracted or *covertly* shifted to internal thoughts. We aim to identify shifts of attention from the visual task (i.e. external stimuli) to internal thought.

Christoff et al. proposed a conceptual space of internal thoughts that includes goal-directed thought, creative thinking, mind wandering, dreaming and rumination [5]. The most well-studied topic is mind wandering detection, generally defined as task-unrelated thought [35] or perceptually decoupled thought [31]. However, knowing when users have dived into a task-related thought and stopped receiving external messages is similarly important for an intelligent system to provide effective interventions. Therefore, our goal is to detect attention shift from the visual material to the general scope of internal thoughts.

There are two points worth noting: (1) externally and internally directed thoughts often co-occur [9][34]. As such, we aim to only detect the period when the focus of attention is shifted from the visual task to internal thought. (2) Attention shift in this paper refers to the shift from external visual task to internal thought. It is different from the *covert attention* shift [10][36], which refers to a spatial shift of visual attention. Previous studies have relied on long-term observations of various eye movement characteristics to infer users' attention [11][17][18], however, such characteristics often have inextricable links with multiple factors, including human intention and interface layout. In contrast, we believe eye vergence behavior can be more indicative of attention and robust to biases. Therefore, we exploit eye vergence for moment-to-moment detection of internal thought.

4 Automatic Detection of Internal Thought

Vergence refers to movements where the eyes move in opposite directions [39]. Focusing on a close object makes the eyes rotate towards each other (converge), while focusing on a distant object makes the eyes rotate away from each other (diverge).

We develop two groups of vergence features: The first is extracted from pairs of left and right eye gaze estimates, while the second is extracted from fixations. The pair-based vergence features are extracted by identifying all valid gaze point pairs in the time window, and then calculating the mean and standard deviation (SD) of their distances and angles.

To calculate the fixation-based vergence features, we first resample eye tracking data at a fixed frame rate (60FPS) and filter it [4]. We then detect fixations using dispersion-threshold identification (I-DT) algorithm [28]. We use a duration threshold of 80ms [15] and a dispersion threshold of 80 pixels (≈1° in our experimental setting). We measure the disparity and angle between the centroids of the fixation points for the left and right eye. Similarly, we extract the minimal bounding circles for the corresponding gaze points and calculate their center distance and angle. Additionally, we included the center distance normalized by the sum of the radii of the bounding circles.

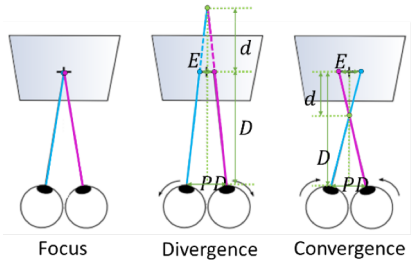


Figure 3. A simplified model for displacement estimation between visual focus and screen surface.

Although the disparity in screen coordinates between the gaze estimates of the left and right eyes provides a direct vergence representation, the distance to the screen can delineate the visual focus more precisely. We use a simplified model (see Figure 3) extended from Kudo et al.'s work [20] to approximate the visual focus displacement from the screen as

$$d = \begin{cases} E \cdot D/(PD - E), \text{divergence} \\ E \cdot D/(PD + E), \text{convergence} \end{cases}$$

where D is eye-to-screen distance, PD denotes pupillary distance and $E = \beta \|\boldsymbol{g}^L - \boldsymbol{g}^R\|_2$ indicates gaze disparity in the world coordinate, which is transformed from gaze disparity $\|\boldsymbol{g}^L - \boldsymbol{g}^R\|_2$ in the screen coordinate to the world coordinate by a constant β (=0.283). Please note that d is positive in divergence and negative in convergence.

Table 1 summarizes all features used in this work, including the proposed vergence features and those used in prior works. In total, we extracted 120 features. The shaded features (classic feature set) are adapted from previous works and mainly cover fixations, saccades, and blinks [2][11][17][18][23]. Following their methods, the bold attributes (i.e. fixation duration, saccade duration, length and velocity, angle between saccades) use multiple descriptive statistics as features, including mean, standard deviation, median, min, max, range, kurtosis, and skewness. Different from studies that were only interested in the overall movement of both eyes, we extract saccadic features (duration, length, velocity, and angles) separately for each eye to capture different oculomotor behaviors between left and right eye. In addition, as our goal is moment-to-moment detection with a window no longer than 1s, most of our new features use only mean and SD. Finally, since we aim for a user-independent model that needs to generalize to completely unknown users, within-participant feature normalization is not performed.

Type	Feature description
Vergence and distance (17)	Disparity of gaze point pairs (mean, SD); gaze focus distance from screen (mean, SD)
	Disparities of gaze point sets – centroid distance; center distance of minimal bounding circles; and center distance over sum of radii
	Direction/angle of gaze point pairs (mean, SD); and their mean centroid and center angles
	Distance between eyes and screen; pupillary distance (mean, SD)
Fixation (13)	Radii of minimal bounding circles
	Duration of fixation; fixation duration and total number over the window
	Duration ratio of fixation over saccade
Saccade (86)	**Duration, length, and velocity of Saccades**; duration and total number over the window
	Angles in degrees between saccades relative to the x-axis and to the previous saccade
	Proportion of horizontal saccades with angles of [-30º, 30º] relative to the x-axis
Blink (4)	Duration of blink (mean, SD); blink duration and total number over the window

The bold features use multiple descriptive statistics, including mean, standard deviation (SD), median, min, max, range, kurtosis, and skewness. Numbers in brackets indicate the number of features.

Table 1. Gaze features used in our evaluations. Shaded cells are features described in prior works [2][11][17][18][23].

5 Annotating Fine-Grained Internal Thought

Ground truth of internal thought is challenging to obtain. A related line of work is mind wandering annotation. Existing experimental paradigms are either probe-caught [2][17][18][27][40] or self-caught [11][23][27][34]. *Probe-caught* paradigms interrupt participants intermittently during or upon the completion of a task to acquire their experience. However, it can be ambiguous whether a mind wandering episode that ends shortly before the probe should be counted or not; participants' sense of time can also affect self-report reliability. Besides, it is almost impossible for existing paradigms to mark the start and end of mind wandering. That is, even if the participant is accurately aware of his/her own mind wandering, existing paradigms can only make a coarse-grained annotation as to whether mind wandering occurred during the time window.

In contrast, *self-caught* paradigms require participants to consciously reflect upon their experience. This alleviates the annotation precision problem, as it gives an approximation of the end. However, while mind is wandering, people react relatively slowly [22][40], and there is a high delay uncertainty between the report time and the end of the mind wandering episode. Moreover, people often fail to notice their mind wandering at all [34]. Likewise, probe-caught and self-caught annotations for internal thought suffer from the same problems.

5.1 Experimental Paradigm for Annotation

To address some of these limitations, we propose a new experimental paradigm for fine-grained internal thought annotation. This paradigm allows us to assess internal thought behavior in a more fine-grained manner and yields clean data for training internal thought detection methods. The key idea is to blur the visual stimulus (e.g. a video, document or webpage) gradually and periodically at random intervals of 10-20s (see Figure 4). Participants were instructed to click or press a key once they notice the blur effect, which immediately deblurs the stimulus. We specifically opted for gradual blurring because a

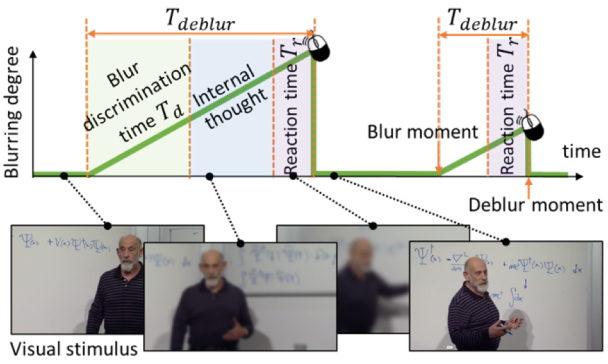


Figure 4. Experimental paradigm for fine-grained internal thought annotation. Participants watched a video which blurs gradually at random intervals. They were instructed to click and deblur the video once they noticed the blur effect. Being visually on-task leads to a quick deblur. In contrast, attending to internal thought may cause ignorance of the blur effect and thus yield a slow deblur. Excluding the time for visual discrimination and reaction gives a conservative internal thought duration.

sudden blurring can be visually salient and disrupt an ongoing internal thought. To generate the blur effect, we used a Gaussian kernel with an aperture size of 15 pixels and with the Gaussian standard deviation, $sigma$, defined by a linear function of the time in seconds, t, elapsed from the blur beginning, i.e. $sigma = \alpha \cdot t$, where α is a factor of blurring speed.

5.2 Identifying Start and End of Internal Thought

To identify the start and end of internal thought, we aimed to investigate two hypotheses:

H1: Users who are visually on-task can perceive the gradual blur effect consistently when it reaches the degree (i.e. $sigma = \alpha \cdot T_d$) that corresponds to the discrimination time T_d.

H2: After attention shifts from the visual task to internal thought, users may fail to notice the blur effect. Therefore, their time to deblur, T_{deblur}, is slower and greater than T_d.

The key idea is that there is a discrimination time, T_d, by which the blur effect should be so obvious that users who are visually on-task can notice it immediately, and thus click to deblur visual stimulus. Since deblur action only happens if the participant is visually on-task, it marks the end of internal thought. While $T_{deblur} \leq T_d$ may be mainly due to the inability of the eye to discriminate small amounts of blur, $T_{deblur} > T_d$ may imply attention shift to internal thought. In other words, T_d provides a conservative start of internal thought. Participants' reaction time is also factored into this paradigm. Based on the human reaction speed reported in prior studies [22][40], a small response time T_r=0.3s is included just before the deblur action is triggered (see Figure 4). Given this paradigm, the period of internal thought is annotated from T_d after the blur effect begins, to T_r before the deblur moment as shown in Figure 4. Our paradigm is therefore able to provide a conservative estimate for internal thought at a fine-grained resolution.

6 Understanding Perception of Gradual Blurring

We first investigated hypothesis *H1*. We studied the dominant factor for blur perception and deblur action to determine the discrimination time T_d. Throughout our experiments, we used lecture video as visual material, as online lecture viewing can be a good use case for internal thought detection [18][26].

6.1 Experiment Setting

Human blur perception is affected by multiple factors, including scene motion, luminance, depth, and screen attributes [42]. Since screen attributes in real use are a factor we can hardly control and the depth of video stimuli on screen is rather consistent, we focus on remaining factors, i.e. blurring speed, motion, and luminance.

For simplicity, we studied deblur actions under three conditions for each factor. Specifically, we evaluated three levels of blurring speed: fast ($\alpha = 2$), medium ($\alpha = 1$), and slow ($\alpha = 0.5$). We extracted three scene motion levels from our video stimuli: small (only lecture slides were shown), medium (a small window in the corner of the screen showed the lecturer; the rest of the screen showed the slides), and large (the screen showed the lecturer, as in Figure 4). We also adjusted the pixel intensity of the videos to obtain three levels of luminance: dark (-50), normal (no change), and bright (+50).

We recruited 11 participants for this blur perception experiment (seven female; mean age: 27.2, SD: 4.4). Participants have seated about 60 cm away from the monitor in a comfortable posture without using a chin rest. We used a 22" monitor at 1680 ×1050 resolution to display the stimuli in full-screen mode and a Tobii EyeX remote eye tracker recording binocular gaze at around 60 FPS. The eye tracker was carefully calibrated before each session to ensure binocular tracking performance. The accuracy of this eye tracker is around 0.6° and 0.9° in the x- and y-direction, respectively, and its precisions in both direction are around 0.9° [12]. Participants were instructed to stay focused and click to deblur the video as soon as they perceived the blur effect, which was generated randomly after 2-5s from the beginning of each video clip. In total, we prepared 22 video clips (eight of small motion, six of medium, and eight of large), each of which lasted for 10s. With the combinations of different luminance degrees and blurring speeds, this experiment took 33 min per participant.

6.2 Validating *H1* about Discrimination Time

We plot the probability distributions of deblur actions of all the participants under different conditions of blurring speed, motion, and luminance. Figure 5 shows their conditional probability mass functions. As expected, blurring speed highly affects the deblur action. In general, fast blurring speed results in a quick deblur. Slow speed leads to fatter tails and greater variability, while the

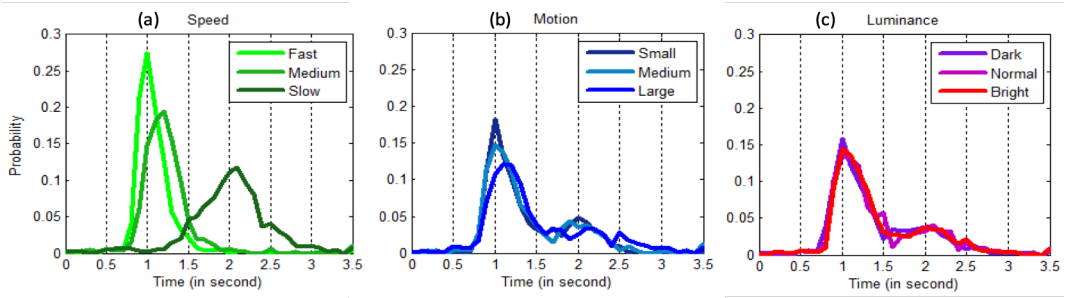


Figure 5. Probability mass functions of deblur actions under different (a) blurring speeds, (b) scene motions, and (c) luminance degrees. The scene motion and luminance do not affect the deblur action much. At a medium blurring speed, a clear majority of participants perceived the blur effect and triggered the deblur action within 1.5s (*H1*).

medium and fast have much narrower shapes. The deblur probabilities of fast and medium speeds peak at 1s and 1.2s, and the clear majority of the actions were completed before 1.5s. Based on this finding, we used the medium blurring speed ($\alpha = 1$) in our following experiment because it produced a consistent deblur pattern with acceptable variability across participants and posed less salient visual impact than fast blurring speed.

It is interesting to see that scene motion and luminance do not obviously affect deblur actions. The curves in Figure 5 (b) and (c) significantly overlap. This further indicates that it is very likely that people who are visually on-task can perceive the blur effect and trigger deblur action within 1.5s under medium blurring speed. Excluding the reaction time T_r (=0.3s) gives us T_d=1.2s.

This experiment corroborates *H1*: that users visually on-task can perceive the blur effect within a consistent time. It also provides evidence to determine the discrimination time T_d, which relates to the start and end of internal thought.

7 Detecting Internal Thought in Video Viewing

This section presents our investigation on hypothesis *H2* and then the classification performance of internal thought.

7.1 Experiment Procedure

A common lecture timespan is around 50-60 min. Therefore, we prepared six video clips for this experiment, each of which lasted for 10 min. As discussed previously, different types of scenes do not affect blur perception much. We thus simplified the scene types. Three video clips contained only the slides view, while the other three showed the view in lecture room. To increase the chance of capturing sufficient data of spontaneous internal thought, our videos were chosen to be challenging to follow: the topics were difficult, such as Riemann Hypothesis and Fermat's Last Theorem, and we also slowed down the video clips to 85% of the original speed. Our post-experiment interview confirmed that participants indeed found difficulty to visually attend to these videos and their minds wandered frequently.

We recruited 24 participants (12 female; mean age: 25.6, SD: 3.1) for this experiment, and used the same experiment setting as previous experiment. We recorded screen video as well as facial video using a webcam mounted on the monitor. Due to the decrease of eye tracking accuracy near screen edges [12], we placed the video stimulus in the center of the screen with a size of 18.1 cm by 10.2 cm, which is the default size of YouTube video in our setting. Participants were allowed to sit freely but recommended to avoid significant head movements.

7.2 Validating *H2* about Actual Deblur Time

Interestingly, there were many slow deblur actions (T_{deblur}>T_d =1.5s) during video viewing. This verifies our hypothesis *H2* that people who attend to internal thought fail to notice the blur effect and delay the deblur. Figure 6 presents the histogram of deblur time T_{deblur} across participants. We see that T_{deblur} has a bimodal distribution, which is illustrated by the orange and green dashed lines. There is a clear peak near 1s. This is in accordance with the conditional probability mass functions (see Figure 5) when participants were visually on-task. The other peak is at around 2.5s, which is likely to reflect the deblur distribution during internal thought. Since these two peaks are close to each other and suggest two highly overlapping distributions, to obtain reliable annotated data, we only assume that a long T_{deblur} is indicative of internal thought but not the other way around.

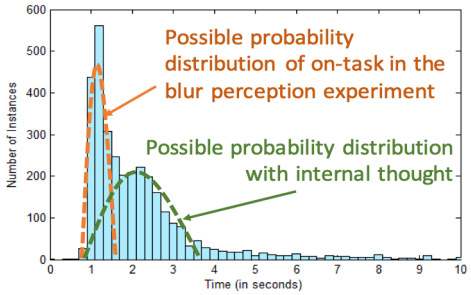
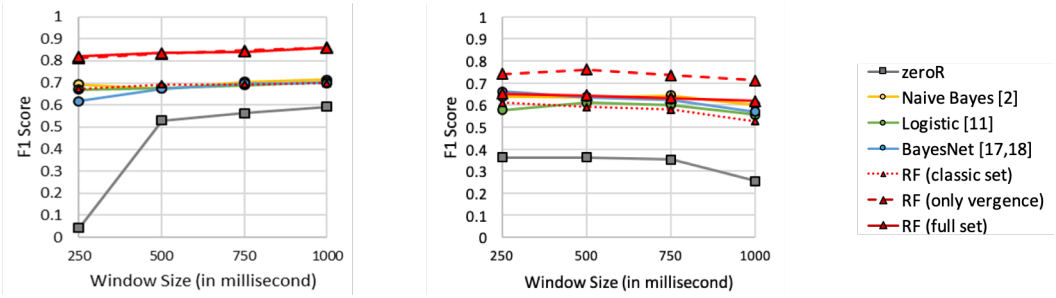


Figure 6. Histogram of deblur time T_{deblur} from the start of each blur effect. The bimodal distribution indicates that internal thoughts delay deblur actions of participants (*H2*).



(a) Internal thought vs deliberate on-task (b) Internal thought vs spontaneous on-task

Figure 7. Weighted average F1 scores of the classifications (a) between internal thought and deliberate on-task and (b) between internal thought and spontaneous on-task. The proposed method (red dashed line) with vergence features alone can outperform the state of the arts and achieve moment-to-moment detection with eye information no more than 1s.

7.3 Annotating the Video Viewing Dataset

We define three types of mental state in this study: (1) internal thought, (2) deliberate on-task, and (3) spontaneous on-task. The classification between internal thought and spontaneous on-task in real-use situation is our ultimate goal, however, our results show that collecting deliberate on-task data is beneficial for training.

7.3.1 Annotating internal thought data. Applying our internal thought annotation paradigm produced 2943s of "internal thought" data across participants in our video viewing dataset.

7.3.2 Annotating deliberate on-task data. We recruited another 16 participants (six female; mean age: 26.9, SD:2.0) and instructed them to view shorter editions of the six video clips (30s each) in complete focused attention. This produced 2880s of clean training data of "*deliberate* on-task" across participants. Although the deliberate data might sound artificial, it is an effective way to obtain clean and reliable visually on-task data and it alleviates the data skew issue in training, due to the scarcity of spontaneous on-task data in our nonintrusive setting.

7.3.3 Annotating spontaneous on-task data. We aim to obtain spontaneous on-task data in a nonintrusive manner. We therefore postulate that participants can maintain on-task for at least 1.5s after they refocus back on the visual task. This should be a plausible assumption, since the blur perception experiment suggests that participants can sustain attention at least for a discrimination time T_d (=1.5s). Besides, a close scrutiny of the facial video discloses that participants oftentimes were not fully attentive after a slow deblur action. Instead, they were in a state that is not completely zoned out but on the verge of focused attention. To reduce the ambiguity of evaluation data, we annotate *the 1.5s period starting from a reaction time, T_r, before each fast deblur ($T_{deblur} \leq 1.5s$) action as spontaneous on-task*. This produced 1796s of "*spontaneous* on-task" data across participants.

7.4 Results of Internal Thought Detection

We now present a quantitative evaluation of internal thought detection. To evaluate the effectiveness of moment-to-moment detection, we used sliding windows within 1s (250, 500, 750, and 1000ms) to generate instances for training and testing. The resultant datasets ranged between 60000 (250ms) and 8500 (1000ms) instances. We conducted a leave-one-participant-out cross-validation comparing with state of the arts [2][11][17][18].

Given the similar nature of task, we compared our method with three state-of-the-art methods for mind wandering detection during reading [2][11], lecture viewing [17], and interacting with a tutorial system [18]. These methods use the shaded features in Table 1, though previous work did not use blink features in non-reading tasks [17][18]. The following results are labeled by the classifier used to generate them: Naïve Bayes [2], Logistic [11], and Bayes Network [17][18]. As a baseline, we also included ZeroR, which always predicts the major class in training set.

7.4.1 Classification between internal thought and deliberate on-task. We first investigate whether our classifier can distinguish between internal thought and deliberate on-task. Figure 7 (a) shows the F1 score of Random Forest (RF) classifier against state of the arts and baseline across different window sizes. Since ZeroR heavily depends on class distribution in training set, it can be susceptible to window size, which affects the number of instances. For example, the 250ms window leads to a very low F1 (0.04) for ZeroR (see Figure 7 (a)). However, it is encouraging that all other methods are more robust to the data distribution and achieves significantly higher F1 scores than baseline ($p<0.001$).

It is even more encouraging that only using vergence features performs as well as using the full feature set for RF both reach an average F1 score of 0.84 across different windows. This is significantly higher performance than that achieved by the state of the art (F1 around 0.68, $p<0.001$). Although performance tends to increase as window size increases, the difference is not significant ($p=0.09$ between 250ms and 1000ms).

These results also imply that the classic features (fixations, saccades, and blinks) within 1s windows do not contribute much to performance, at least for video viewing task. Fixations, saccades and blinks in video viewing can be affected by numerous factors, such as video saliency and user intention. This suggests

that in natural environments using vergence features for moment-to-moment detection is preferable and even more appropriate.

7.4.2 Classification between internal thought and spontaneous on-task. The vergence features appear effective in distinguishing internal thought from deliberate on-task, however, an even more interesting question is: can the deliberate on-task data contribute to the discrimination between internal thought and spontaneous on-task?

Similar to the previous evaluations, we performed a leave-one-participant-out cross-validation. Specifically, on each iteration, we trained on the internal thought and spontaneous on-task data of the training participants as well as the deliberate on-task data from the separate group of participants, and tested on the left-out participant. Figure 7 (b) presents the F1 score of RF classifier with the vergence features against the baseline and state of the arts in discriminating spontaneous on-task and internal thought.

It is encouraging that RF (only vergence) yields an obviously higher performance than its counterparts. It achieves an average F1 of 0.74, while the results of the other methods range from 0.52 to 0.66. That is, a 12.1% improvement over the state of the art. Surprisingly, RF (full set) fails to perform as well as RF (only vergence) in this case. This is probably because the more natural contexts in this experiment made classic features more susceptible to noise. Having said that, RF (full set) with the vergence features still improves significantly over the baseline ($p \leq 0.008$). We also see that time windows do not affect the performance much. In general, this indicates that vergence features can effectively distinguish internal thought from spontaneous on-task. Additionally, this result also suggests the effectiveness of training with deliberate on-task data.

8 Discussion

In this work, we are the first to exploit eye vergence behavior for light-weight and moment-to-moment detection of internal thought. This technique shows promising results for internal thought detection during video viewing. Our evaluations demonstrated that our method using vergence feature alone can outperform existing method using fixations, saccades, and blinks.

To acquire fine-grained annotation of internal thought and meanwhile maintain natural user behavior in task, we propose a novel experimental paradigm for annotation, which exploits human blur perception. We also conducted a human blur perception study to determine the related parameters. Compared with traditional self-caught and probe-caught techniques, our annotation paradigm estimates a conservative period of internal thought and it requires less intrusive user intervention, allowing users to sustain attention to the main task and behave more naturally. It therefore opens new opportunities for human behavioral and psychological studies on attention.

One important property that differentiates our method from the state of the art is that vergence features can be extracted from a very small time window (1s). This allows for moment-to-moment detection of internal thought and thus enables fine-grained understanding of users' attentive state. Taken together with the context of visual material, future intelligent interfaces can provide effective interventions to optimize memory retention, knowledge acquisition, and recommendation. With the rapid emergence of multimedia interfaces and increasing demand of understanding and adapting to individual users, our method provides a light-weight solution to moment-to-moment detection of internal thought and thus paves the way for attention-aware applications.

Despite these promising findings, there are still some limitations that we aim to address in future work. First, the eye tracker (Tobii EyeX) we used in our study is a consumer-grade tracker the same as Hutt et al. [18]. This is to study the feasibility of real-time attention-aware interfaces that use only affordable devices in real-world environments. This attempt thus alleviates the constraint of the expensive research-grade eye tracker as pointed out by Bixler and D'Mello [2]. Although EyeX gives a similar accuracy (offset from the true gaze point) as Tobii Pro X2-60 as well as a reasonable precision [12] and its firmware is optimized for real-time interaction [16], its frame rate still limits the accurate detection of some saccadic behaviors. This might affect our comparison of using the "classic" eye movement features, but it is currently unavoidable as research-grade eye trackers are too expensive for home-use intelligent interfaces. Besides, since we posed no explicit constraint on the head movement in our experiment to maintain the natural user behavior, this might lead to an accuracy drop of eye tracking [24].

In future, we plan to study vergence behavior by representation learning from large-scale in-the-wild data. As Solé Puig et al. pointed out, vergence can be affected by covert spatial attention [36] and we foresee that accounting for contextual information [3] or activities information [38] can result in a better representation for internal thought detection. Besides, exploring other visual effects in addition to blur effect for internal thought annotation paradigm can also be interesting especially for in-situ interactions.

9 Conclusion

We proposed a method for user-independent, moment-to-moment detection of internal thought from eye vergence behavior. It can detect attention shift from visual task to internal thought, and the other way around. We further proposed a novel experimental paradigm to identify the start and end of internal thought in a fine-grained manner. Our results demonstrated the effectiveness of our method for moment-to-moment detection of internal thought. As moment-to-moment detection allows for a fine-grained understanding of human mental activities, our findings not only contribute to the fundamental study of human mental states but also open up new applications, e.g. in mental health. As such, we believe moment-to-moment detection of internal thought can be an indispensable component for user experience research.

ACKNOWLEDGMENTS

This work was funded in part by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University, Germany.

REFERENCES

[1] Benedek, M. et al. 2017. Eye Behavior Associated with Internally versus Externally Directed Cognition. Frontiers in Psychology. 8, June (Jun. 2017), 1–9. DOI:https://doi.org/10.3389/fpsyg.2017.01092.

[2] Bixler, R. and D'Mello, S. 2016. Automatic gaze-based user-independent detection of mind wandering during computerized reading. User Modeling and User-Adapted Interaction. 26, 1 (Mar. 2016), 33–68. DOI:https://doi.org/10.1007/s11257-015-9167-1.

[3] Bulling, A. and Zander, T.O. 2014. Cognition-Aware Computing. IEEE Pervasive Computing. 13, 3 (Jul. 2014), 80–83. DOI:https://doi.org/10.1109/MPRV.2014.42.

[4] Casiez, G. et al. 2012. 1 € filter: a simple speed-based low-pass filter for noisy input in interactive systems. Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12 (New York, New York, USA, 2012), 2527.

[5] Christoff, K. et al. 2016. Mind-wandering as spontaneous thought: a dynamic framework. Nature Reviews Neuroscience. 17, 11 (Nov. 2016), 718–731. DOI:https://doi.org/10.1038/nrn.2016.113.

[6] Conati, C. et al. 2013. Eye-Tracking for Student Modelling in Intelligent Tutoring Systems. Design Recommendations for Intelligent Tutoring Systems. 227–236.

[7] D'Mello, S.K. 2016. Giving Eyesight to the Blind: Towards Attention-Aware AIED. International Journal of Artificial Intelligence in Education. 26, 2 (Jun. 2016), 645–659. DOI:https://doi.org/10.1007/s40593-016-0104-1.

[8] Dingler, T. 2016. Cognition-Aware systems as mobile personal assistants. UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. (2016), 1035–1040. DOI:https://doi.org/10.1145/2968219.2968565.

[9] Dixon, M.L. et al. 2014. A framework for understanding the relationship between externally and internally directed cognition. Neuropsychologia. 62, (Sep. 2014), 321–330. DOI:https://doi.org/10.1016/j.neuropsychologia.2014.05.024.

[10] Engbert, R. and Kliegl, R. 2003. Microsaccades uncover the orientation of covert attention. Vision Research. 43, 9 (2003), 1035–1045. DOI:https://doi.org/10.1016/S0042-6989(03)00084-1.

[11] Faber, M. et al. 2017. An automated behavioral measure of mind wandering during computerized reading. Behavior Research Methods. (2017). DOI:https://doi.org/10.3758/s13428-017-0857-y.

[12] Feit, A.M. et al. 2017. Toward Everyday Gaze Input: Accuracy and Precision of Eye Tracking and Implications for Design. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17 (New York, New York, USA, 2017), 1118–1130.

[13] Franklin, M.S. et al. 2011. Catching the mind in flight: Using behavioral indices to detect mindless reading in real time. Psychonomic Bulletin and Review. 18, 5 (2011), 992–997. DOI:https://doi.org/10.3758/s13423-011-0109-6.

[14] Franklin, M.S. et al. 2013. Window to the wandering mind: Pupillometry of spontaneous thought while reading. The Quarterly Journal of Experimental Psychology. 66, 12 (Dec. 2013), 2289–2294. DOI:https://doi.org/10.1080/17470218.2013.858170.

[15] Hansen, D.W. and Ji, Q. 2010. In the eye of the beholder: a survey of models for eyes and gaze. IEEE transactions on pattern analysis and machine intelligence. 32, 3 (2010), 478–500. DOI:https://doi.org/10.1109/TPAMI.2009.30.

[16] http://developer.tobii.com/: 2018. .

[17] Hutt, S. et al. 2017. Gaze-based Detection of Mind Wandering during Lecture Viewing. 10th International Conference on Educational Data Mining (2017).

[18] Hutt, S. et al. 2016. The Eyes Have It: Gaze-based Detection of Mind Wandering during Learning with an Intelligent Tutoring System. Proceedings of the 9th International Conference on Educational Data Mining, International Educational Data Mining Society (2016), 86–93.

[19] Killingsworth, M.A. and Gilbert, D.T. 2010. A Wandering Mind Is an Unhappy Mind. Science. 330, 6006 (Nov. 2010), 932–932. DOI:https://doi.org/10.1126/science.1192439.

[20] Kudo, S. et al. 2013. Input method using divergence eye movement. CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13 (New York, New York, USA, 2013), 1335.

[21] Li, J. et al. 2016. Your Eye Tells How Well You Comprehend. IEEE 40th Annual Computer Software and Applications Conference (COMPSAC) (2016), 503–508.

[22] Mijović, P. et al. 2017. Towards continuous and real-time attention monitoring at work: reaction time versus brain response. Ergonomics. 60, 2 (Feb. 2017), 241–254. DOI:https://doi.org/10.1080/00140139.2016.1142121.

[23] Mills, C. et al. 2016. Automatic Gaze-Based Detection of Mind Wandering during Narrative Film Comprehension. Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016) (2016), 30–37.

[24] Niehorster, D.C. et al. 2017. What to expect from your remote eye-tracker when participants are unrestrained. Behavior Research Methods. (Feb. 2017). DOI:https://doi.org/10.3758/s13428-017-0863-0.

[25] Olney, A.M. et al. 2015. Attention in Educational Contexts: The Role of the Learning Task in Guiding Attention. The Handbook of Attention. MIT Press. 623–642.

[26] Pham, P. and Wang, J. 2015. AttentiveLearner: Improving Mobile MOOC Learning via Implicit Heart Rate Tracking. International Conference on Artificial Intelligence in Education (2015), 367–376.

[27] Reichle, E.D. et al. 2010. Eye Movements During Mindless Reading. Psychological Science. 21, (2010), 1300–1310. DOI:https://doi.org/10.1177/0956797610378686.

[28] Salvucci, D.D. and Goldberg, J.H. 2000. Identifying fixations and saccades in eye-tracking protocols. Proceedings of the symposium on Eye tracking research & applications - ETRA '00 (New York, New York, USA, New York, USA, 2000), 71–78.

[29] Sattar, H. et al. 2017. Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) (Oct. 2017), 2740–2748.

[30] Sattar, H. et al. 2015. Prediction of search targets from fixations in open-world settings. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun. 2015), 981–990.

[31] Schooler, J.W. et al. 2011. Meta-awareness, perceptual decoupling and the wandering mind. Trends in Cognitive Sciences. (Jun. 2011). DOI:https://doi.org/10.1016/j.tics.2011.05.006.

[32] Smallwood, J. et al. 2004. Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. Consciousness and Cognition. 13, 4 (Dec. 2004), 657–690. DOI:https://doi.org/10.1016/j.concog.2004.06.003.

[33] Smallwood, J. et al. 2004. The consequences of encoding information on the maintenance of internally generated images and thoughts: The role of meaning complexes. Consciousness and Cognition. 13, 4 (Dec. 2004), 789–820. DOI:https://doi.org/10.1016/j.concog.2004.07.004.

[34] Smallwood, J. and Schooler, J.W. 2006. The restless mind. Psychological Bulletin. 132, 6 (2006), 946–958. DOI:https://doi.org/10.1037/0033-2909.132.6.946.

[35] Smallwood, J. and Schooler, J.W. 2015. The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness. Annual Review of Psychology. 66, 1 (2015), 487–518. DOI:https://doi.org/10.1146/annurev-psych-010814-015331.

[36] Solé Puig, M. et al. 2013. A Role of Eye Vergence in Covert Attention. PLoS ONE. 8, 1 (Jan. 2013), e52955. DOI:https://doi.org/10.1371/journal.pone.0052955.

[37] Solé Puig, M. et al. 2015. Attention-Related Eye Vergence Measured in Children with Attention Deficit Hyperactivity Disorder. PLOS ONE. 10, 12 (Dec. 2015), e0145281. DOI:https://doi.org/10.1371/journal.pone.0145281.

[38] Steil, J. and Bulling, A. 2015. Discovery of everyday human activities from long-term visual behaviour using topic models. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15 (New York, New York, USA, 2015), 75–85.

[39] Toates, F.M. 1974. Vergence eye movements. Documenta Ophthalmologica. 37, 1 (1974), 153–214. DOI:https://doi.org/10.1007/BF00149678.

[40] Unsworth, N. and Robison, M.K. 2016. Pupillary correlates of lapses of sustained attention. Cognitive, Affective, & Behavioral Neuroscience. April (2016), 601–615. DOI:https://doi.org/10.3758/s13415-016-0417-4.

[41] Walcher, S. et al. 2017. Looking for ideas: Eye behavior during goal-directed internally focused cognition. Consciousness and Cognition. 53, (Aug. 2017), 165–175. DOI:https://doi.org/10.1016/j.concog.2017.06.009.

[42] Watson, A.B. and Ahumada, A.J. 2011. Blur clarified: A review and synthesis of blur discrimination. Journal of Vision. 11, 5 (Sep. 2011), 10–10. DOI:https://doi.org/10.1167/11.5.10.

[43] Xiao, X. and Wang, J. 2017. Undertanding and Detecting Divided Attention in Mobile MOOC Learning. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17. (2017), 2411–2415. DOI:https://doi.org/10.1145/3025453.3025552.