Are you with me? Measurement of Learners' Video-Watching Attention with Eye Tracking

Namrata Srivastava
The University of Melbourne, and
Monash University
Melbourne, VIC, Australia
srivastavan@student.unimelb.edu.au

Sadia Nawaz
The University of Melbourne
Melbourne, VIC, Australia
nawazs@student.unimelb.edu.au

Joshua Newn
The University of Melbourne
Melbourne, VIC, Australia
joshua.newn@unimelb.edu.au

Jason Lodge
The University of Queensland
Brisbane, QLD, Australia
jason.lodge@uq.edu.au

Eduardo Velloso
The University of Melbourne
Melbourne, VIC, Australia
eduardo.velloso@unimelb.edu.au

Sarah Erfani
The University of Melbourne
Melbourne, VIC, Australia
sarah.erfani@unimelb.edu.au

Dragan Gašević
Monash University
Melbourne, VIC, Australia
dragan.gasevic@monash.edu

James Bailey
The University of Melbourne
Melbourne, VIC, Australia
baileyj@unimelb.edu.au

ABSTRACT

Video has become an essential medium for learning. However, there are challenges when using traditional methods to measure how learners attend to lecture videos in video learning analytics, such as difficulty in capturing learners' attention at a fine-grained level. Therefore, in this paper, we propose a gaze-based metric—"with-me-ness direction" that can measure how learners' gaze-direction changes when they listen to the instructor's dialogues in a video-lecture. We analyze the gaze data of 45 participants as they watched a video lecture and measured both the sequences of with-me-ness direction and proportion of time a participant spent looking in each direction throughout the lecture at different levels. We found that although the majority of the time participants followed the instructor's dialogues, their behaviour of *looking-ahead*, *looking-behind* or *looking-outside* differed by their prior knowledge. These findings open the possibility of using eye-tracking to measure learners' video-watching attention patterns and examine factors that can influence their attention, thereby helping instructors to design effective learning materials.

CCS CONCEPTS

• **Applied computing → E-learning**.

KEYWORDS

learning analytics, eye-tracking, gaze direction, video lecture, co-attention

LAK21, April 12–16, 2021, Irvine, CA, USA
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448148

ACM Reference Format:
Namrata Srivastava, Sadia Nawaz, Joshua Newn, Jason Lodge, Eduardo Velloso, Sarah Erfani, Dragan Gašević, and James Bailey. 2021. Are you with me? Measurement of Learners' Video-Watching Attention with Eye Tracking. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3448139.3448148

1 INTRODUCTION

Video lectures have gained prominence as a standard medium for instruction within technology-driven learning environments [19]. However, their growing impact in online learning environments such as massive open learning courses and flipped classrooms offers both new opportunities and challenges to learning at scale [5]. On the one hand, video lectures can assist students in learning complex concepts anytime and anywhere. On the other hand, it is becoming difficult for instructors to understand how students interact with video lectures as there is a lesser opportunity for them to monitor [31]. For example, how likely are the students to follow the instructor's dialogue in a video-lecture? What are the factors that can influence their video-watching behaviours? In this paper, we aim to answer these questions.

Several video learning analytics methods, such as analyzing clickstream interaction patterns of the students (e.g., pause, play, skip, seek-forward, seek-backward) help researchers to understand students' engagement, participation and dropout rates [4, 11, 23]. However, these methods fail to capture the real-intent of students' attention (e.g., play a video but not watch). To gain deeper insights into students' attention patters, researchers have begun exploring the concept of "student-teacher co-attention" in video-lectures using eye-tracking [21, 22]. For instance, Sharma et al. [22] measured the co-attention between instructor's dialogues and students' gaze using a gaze-based metric known as "with-me-ness", and found a positive correlation between with-me-ness and students' learning outcomes. Their findings suggest that with-me-ness can be used as an indicator of students' attention in video-lectures. However, considering that students' attention is not always directed to instructor's dialogue in the video-lecture, e.g., they can look outside the screen while listening to the video lecture, or can fixate to other contents on the screen, an exploration into the *direction* of the student-teacher co-attention would help us to gain better insights into students' video-watching behaviours.

Therefore, in this paper, we propose a new gaze-based metric—**"with-me-ness direction"**— that encapsulates both *temporal* (amount of time) and *directional* aspects (in which particular direction) of student-teacher co-attention. We measured the co-attention between participants' gaze and instructor's dialogues using a dataset consisting of 45 participants' eye-tracking data while they watched a video lecture. We analyzed both the sequences of with-me-ness direction and the proportion of time a participant spent looking in each direction throughout the lecture at different levels. We found that while participants followed the instructor's dialogues in a majority of the time, their behaviour of looking-ahead, looking-behind or looking-outside differed by their prior knowledge.

In the following sections, we present a complete pipeline to extract "with-me-ness direction" features from learners' eye-movements and show their implications for understanding learners' video-watching attention patterns. We also describe the additional preprocessing steps needed to calculate the with-me-ness direction from a video lecture. These additional steps are required due to complexities posed by the dynamic nature of the video-lecture. This paper aims to advance the research in learning analytics using eye-tracking methods by making three contributions. First, we present a novel metric "with-me-ness direction", to measure the video-watching attention patterns between a learner's gaze and the instructor's dialogue in a video lecture. Second, we analyze both the *temporal* and *directional* aspects of the "with-me-ness direction", and show their implications for understanding learners' video-watching behaviours. Third, we demonstrate how learners' viewing strategies can vary with their prior knowledge using the "with-me-ness direction".

2 BACKGROUND

2.1 Video-Based Learning

Video-based learning has become a popular form of learning due to its numerous benefits [19], one of which is that students receive visual and verbal information simultaneously. However, learning theories (such as *dual coding theory* [18] and *cognitive theory of multimedia learning* [15]) suggest there are two cognitive subsystems: one for processing *visual* objects and the other for processing *verbal* objects, and that both are processed separately. Both these channels are distinct in the human mind and can only process a limited amount of information [15]. Due to this limited capacity, learners' have to constantly decide what information they should retain and what to discard while watching video lectures. Nevertheless, presenting visual and verbal information together in a multimedia video can enhance students' conceptual understanding and retention, and may help the students in integrating new knowledge into their existing schema [16]. To date, there is limited research on how students interact with video lectures, and therefore, it is difficult for instructors to determine the effectiveness of the learning design, especially at a fine grain level.

Further, in online learning environments, instructors are usually unaware of the attentional state of the students due to a lack of face-to-face interaction. Several methods using video learning analytics have been proposed by researchers to better understand how students learn with videos. For instance, Kim et al. [11] investigated MOOC learners' in-video dropout rates and interaction peaks by analyzing their video-watching patterns (pausing, playing, replaying and quitting) in online lecture videos. Similarly, Giannakos et al. [4] recorded student clickstream patterns within a video lecture and found that a correspondence exists between students' video navigation patterns (repeated views) and the level of cognition/thinking required for a specific video segment. In another study, Sinha et al. [23] operationalize video lecture clickstreams of students to form cognitive video watching states that summarise students' engagement, their future click interactions and their participation trajectories. Although these clickstream video interaction methods are effective and applicable at large-scale, they do not accurately capture students' attentional state. For example, whether a student is actively paying attention to the video or just playing it in the background while multitasking [11].

2.2 Use of Eye-tracking in Video-Based Learning

Eye-tracking has widely been used in multimedia learning research to gain deeper insights into students' attentional state [1]. Considering the *eye-mind hypothesis* [9], which suggests that a link exists between human-gaze and attention (i.e., we process the information that we visually attend to), researchers have utilized eye-tracking tools to determine what parts of the stimulus a person allocated visual attention, in what order and for how long [28]. Notable examples in multimedia research include investigating cognitive processes (such as selecting, ordering and integrating), the factors that can affect them (such as multimedia content, individual differences, metacognition), and the correlation between the cognitive processes and learning performance (see [1] for a full review). However, the research specifically in video-based learning using eye-tracking methods is currently limited and under-explored.

Prior work in the field of video-based learning using eye-tracking has mostly focused on investigating the effect of instructor's presence in the video-lecture on students' information retention, visual attention, affect or perceived learning [12, 30]. With the growing popularity of low-cost eye trackers and open source eye-tracking software, researchers have utilized eye-tracking methods and techniques for exploring students' cognitive processes during video-watching. For example, Hutt et al. [6] used a low-cost consumer-grade eye tracker to automatically detect mind-wandering states when a student learns from a recorded video. Their study highlighted the importance of attention in learning and how crucial it is to monitor the students' attentional state. Further, the role of attention has been studied extensively in educational research, and recent studies report that inattention has consequences not only for students' retention and learning outcomes [29] but also for their affective experiences [7]. Therefore, building a stronger understanding of the factors that influence attention and investigating new techniques for measuring attention is ever more critical, especially considering the ever-growing dependency and adoption of video lectures as a medium for learning.

2.2.1 Effect of prior knowledge on eye-movements. One factor that can substantially influence students' eye movements during learning is their prior knowledge. The difference in visualization behaviours between experts and novices is largely from long-term memory, experts possess a large number of domain-specific schemas and can bypass working memory capacity limitations; whereas, novices may not have acquired the same relevant schematic knowledge as an expert [10]. Studies in multimedia research have shown that people with higher prior knowledge attend more to relevant information than the people with lower prior knowledge—using both dynamic stimuli (such as fish locomotion video) [8] and static stimuli (such as weather maps) [3]. Lee et al. [13] observed similar results when eye-movements of expert and novice medical professionals were compared in a medical simulation game. They found that learners with high domain-specific prior knowledge (experts) could allocate more attention to task-relevant areas, demonstrate a more systematic approach, and achieve higher performance speed than the learners with low domain-specific knowledge (novices). Recent studies have also explored the differences in novice and expert learners' viewing behaviours within intelligent tutoring systems, including MetaTutor [27] and SQL-Tutor [20]. In both studies with intelligent tutoring systems, it was found that learners exhibit different viewing behaviours based on their prior knowledge. For instance, while learning using MetaTutor, students with high prior content knowledge were looking more frequently on their content-notes than those with lower content knowledge. Similarly, while using SQL-Tutor, advanced learners paid more attention to database schema (AOI) than novices did. While many studies have investigated the effect of prior knowledge on visual attention to date, only a few studies have explored how prior knowledge could influence learners' video watching behaviours. Thus, in this paper, we will explore this effect further.

2.2.2 Measuring co-attention in video-lecture. More recently, the measurement of co-attention between an instructor's dialogues and students' attention in a video-lecture has been explored in the video-based learning literature. In particular, Sharma and colleagues have explored this area of research [21, 22], suggesting that co-attention in a video-lecture could be reflective of students' deeper form of attention. In one study [22], they collected a dataset of 40 students watching two MOOC videos while recording their eye-movement and proposed an eye-tracking based metric 'with-me-ness' that can represent *"To what extent does a learner follow an instructor?"*. They defined with-me-ness at two levels: (1) *perceptual with-me-ness*, which measures how much does a learner follow an instructor's deictic gestures, and (2) *conceptual with-me-ness*, which represents the amount of time a learner followed the instructor's discourse.

In this paper, we extend the idea of conceptual with-me-ness. Sharma et al. [22] computed the conceptual with-me-ness by measuring the proportion of gaze time a learner spent looking at the object verbally referred to by the instructor in the video-lecture (normalized by the slide duration). Although they present a novel approach to measure the co-attention between learners' gaze and the instructor's dialogues, their study lacks sufficient understanding of how learners' gaze-direction changes while listening to the instructor's dialogues. For example, while listening to the lecture, students' gaze can also be directed in a different direction—they can look ahead the active-content, look behind revisiting the already discussed content or look away from the screen.

As determining visual attention allocation can provide more information about learners' cognitive processes [28], we propose a new metric 'with-me-ness direction' in this paper, which can explain both the direction and magnitude of with-me-ness between instructors' discourse and learners' gaze (i.e., how much does a learner follow their instructor, and in which direction). Further, considering the effects of learners' prior knowledge on their viewing behaviour, we aim to answer the following research questions: (1) **RQ1:** What can we infer about learners' video-watching behaviour from the sequences of 'with-me-ness direction'? (2) **RQ2:** Is there a relationship between learners' with-me-ness direction, their learning outcomes, and prior knowledge?

3 DATASET

This section summarizes how we have used the 100-participant dataset collected by Srivastava et al. [24, 26], which focused on understanding participants' learning process in an e-learning environment using unobtrusive sensor technologies. Learners could self-report their perceived difficulty of video lectures in real-time (using a continuous slider) while being unobtrusively recorded by an eye tracker, thermal camera and an RGB webcam in a lab setting. The authors selected video lectures from Lodge et al. [14]'s study as their stimulus, which evaluated different instructional design techniques. The video lectures were on two separate topics — *neuroscience* (explaining the basic working of neurons in the human brain) and *binary numbers* (focusing on binary-number conversion, binary addition and 2's complement). For each topic, there were two variants in terms of presentation format, *text-based* (i.e., bullet point slides with no pictures) and *animation-based* (i.e., animated video with digital ink). All video lectures were synchronized with the instructor's voice.

For our analysis, we use the eye-tracking data from the condition where participants watched the video lecture on binary numbers (text-based variant). The lecture consisted of a slide deck with a voice-over, where each slide contained mostly text and equations, structured into 3-5 bullet points and without any illustrations. The duration of the video lecture was 6 minutes 17 seconds. We selected this video-lecture because it consisted of synchronized audio and text, i.e., the instructor was reading the slide content line by line. The congruence between text and audio in this video lecture provided a strong framework for calculating and analyzing with-me-ness direction features. For clarity, we further describe how the authors collected the dataset and the parts of the dataset we have used in this section.

3.1 Experimental Setup

The multi-sensor experimental setup consisted of a Tobii Pro X2-30 eye-tracker, a Logitech RGB web camera, and an Optris PI-400 thermal camera. The sensors were placed at the top of the desktop monitor and directed at the participants' face to record their eye movements, facial expressions, and facial temperature, respectively. To the right of the setup on the desk, we placed a physical linear

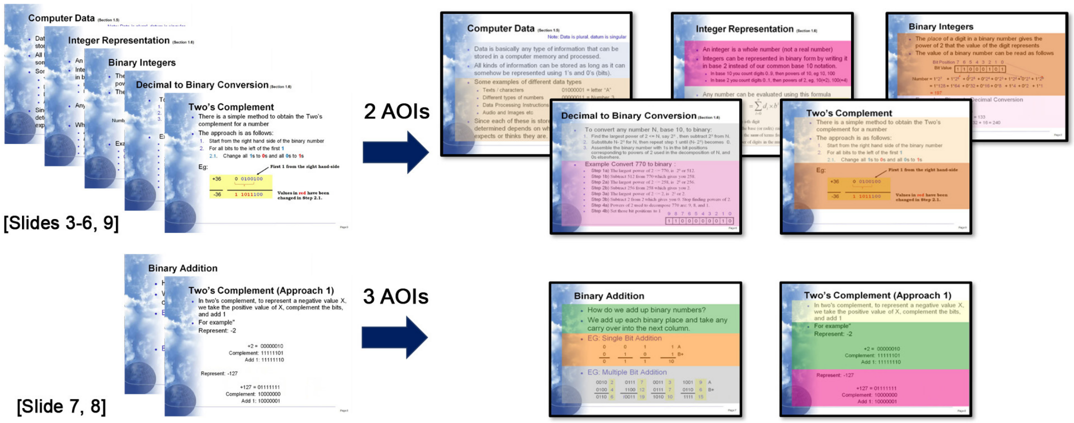


Figure 1: A figure showing how slides were divided into different AOIs. Slide 3-6 and slide 9 were divided into two AOIs (see top), and slide 7 and 8 were divided into three AOIs (see bottom). The different color boxes in the slides represents the AOIs.

slider (Numark Mixtrack PRO Midi Controller) to collect strong ground truth about the difficulty of the video lecture, which the participants held and manipulated throughout the whole lecture (see [26] for details). We connected all the device to a single control station (operated by the experimenter). The data streams were recorded using a custom-built Windows application (using C#) installed on the control station, which facilitated synchronizing the data from all the sensors simultaneously and in real-time using sensors built-in packages.

3.2 Procedure

The main study consisted of two learning tasks. For each learning task, the participants first completed a pre-test, and then watched a video lecture without pausing, rewinding or note-taking. While watching the video lecture, the participants could self-report their perceived difficulty of the lecture by moving the slider continuously. After watching the lecture, they completed a survey questionnaire about the video lecture design (e.g., related to clarity, engagement, and satisfaction), and then completed the post-test. Throughout the learning task, participants' eye-movements, facial expressions, and facial temperature were recorded. The learning tasks differed both by the topic of the lecture, and also the presentation style of the lecture. For example, if a learner watched an "animation-based" neuroscience lecture in the first learning task, then they watched a "text-based" binary lecture in the second learning task, and vice-versa. The order of the topics and the videos styles were counter-balanced between participants.

3.3 Participants

Of the 100 participants in the dataset, only 52 watched the binary lecture in text-format, and we removed 7 participants where the eye-tracking was not correctly recorded (due to loss of gaze data or hardware issues). Hence, our analysis is restricted to 45 participants, aged 19 to 37 (M = 25.02, SD = 4.37). A total of 28 participants identified themselves as female, and 17 as male. In terms of the education level, 19 were undergraduates, 10 were postgraduates (masters), and the remaining 16 were PhD students. Lastly, in terms of the background and knowledge of the learning material, 23 students had previously studied binary numbers, but the remaining 22 students had no prior knowledge of this topic.

3.4 Measures

3.4.1 Prior knowledge. We analyzed the pre-test scores of the participants (conducted before they watched the video lecture). The test contained 9 multiple-choice questions (MCQs) with 4 possible answers, and one "I don't know (IDK)" option. The participants were asked to select the IDK options when they were not sure of the correct answer. The questions were based on three important topics of the lecture: binary-to-decimal conversion (3 questions), binary addition (3 questions), and 2's complement (3 questions). The maximum achievable score was nine by scoring one-point credit for each correct answer. Participants received zero points for incorrect, unselected or IDK options.

3.4.2 Learning outcomes. We analyzed post-test scores of the participants (conducted after they watched the video lecture). The format of the post-test was similar to that of the pre-test containing 9 MCQs with 4 possible answers, and one IDK option. The distribution of the post-test questionnaire was also similar to that of the pre-test MCQs: binary-to-decimal conversion (3 questions), binary addition (3 questions), and 2's complement (3 questions). The maximum achievable score was also nine by scoring one-point credit for each correct answer.

3.4.3 Eye-movements. We utilized the participants' eye-tracking data while they watched the video lecture to evaluate our proposed metric. The eye tracker sampling frequency is 30Hz; however, because of using a custom-built software to synchronize and record all the sensor-data during the experiment, the sampling frequency was reduced to 10Hz.

4 ANALYSIS

This section describes the steps for measuring the co-attention between the instructor's dialogue in the lecture and the learners' gaze direction. As a first step, we processed the video lecture frames to

Table 1: The resulting dataset after transcribing and manually annotating the video lecture

Index	Start (seconds)	End (seconds)	Voice segments	AOI number	Slide number
29	18.75137	19.10204	computer	1	3
30	19.10204	19.55408	data	1	3
31	19.88831	20.14309	is	1	3
...	...	...	...	...	...
441	358.559	360.7069	which is 00100	2	9
442	360.9096	361.4137	we flip	2	9
443	361.5561	363.5453	and we have 11011	2	9

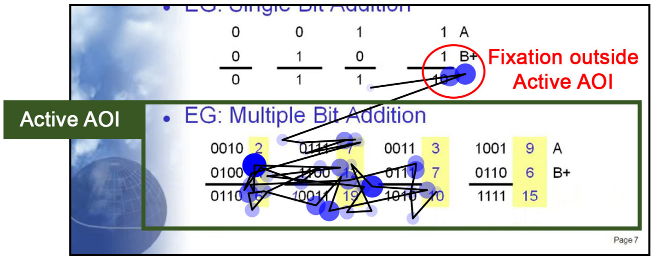


Figure 2: A figure showing the fixations of a student when the instructor was explaining multiple bit addition (Active AOI). The fixations are shown as blue colored circles. Few fixations were observed outside the active AOI.

extract the voice-segments and slides. Then, we defined the areas of interest (AOIs) of each slide and created our lecture annotation dataset. We also pre-processed the raw gaze data into a series of fixations and saccades using a fixation filter. We then extracted five with-me-ness directional features for each AOI. Finally, we present the data analysis procedure by briefly describing the methods used in the paper. The following subsection provides a detailed explanation of each of these steps.

4.1 Pre-processing the Video Lectures

Step 1: Extracting slides and voice-segments from the video lecture*.* The first step for processing the video lecture was to extract the instructor's slides and voice-segments. We extracted the slides from the video frames using OpenCV, which resulted in 10 different video lecture slides. We then manually transcribed the instructor's voice using PRAAT[1]. To map the utterances in the voice-segments to the slides, we further manually annotated the lecture using the Anvil [2] video annotation tool. The resulting voice-segments included the start time and end time of each utterance, the words spoken and the slide number.

Step 2: Segmenting slides into AOIs*.* The next step segments each slide into Areas-of-Interest (AOIs). We were removed slides numbered 1, 2 and 10 from our analysis as they contained only one block of text, explaining the introduction, outline, and summary of the lecture, respectively. The remaining 7 slides had 2 or more AOIs. Most slides have two AOIs, where the first half contained a description of the topic of the slide and the second half displayed examples or formulae (Figure 1-top). As Slides 7 and 8 had two examples each, they were divided into three AOIs, as shown in Figure 1-bottom. All AOIs were manually defined.

Step 3: Synchronising AOIs with the instructor's voice-segments*.* Lastly, we manually mapped the voice-segments to the corresponding AOIs in each slide. The resulting dataset is presented in Table 1. As shown, the dataset contained the start time and the end time of each voice-segment, spoken words, along with slide information such as in which region (AOI) they were present, and on which slide they were referred.

4.2 Pre-processing the Eye-Tracking Data

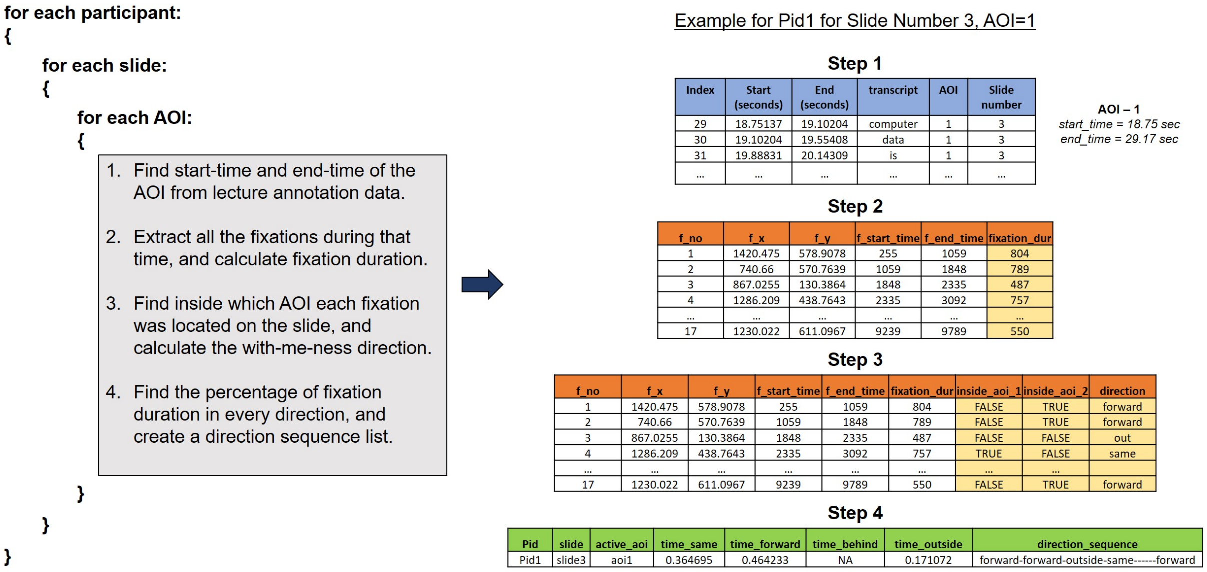
The dataset consisted of timestamped horizontal (x) and vertical (y) gaze coordinates corresponding to where the participant was looking on the screen. To process the raw gaze data into fixations and saccades, we used Olsson's fixation filter [17], setting the peak threshold to 25px and the average window size set to 3 samples (same as [25]).

4.3 Extracting Features from Participants' Eye-Movements

During an initial visual inspection of the distribution of fixations on the slides, we observed that while learners watched a video lecture, the fixations were not only present in the AOIs relevant to the instructor's voice, but also outside of that AOI. For example, Figure 2 shows all fixations from one participant, within a window in which the instructor was describing an example of multiple bit addition. Notice that fixations fall both inside and outside the AOI.

Based on the above observation, we define the area on the slide about which the instructor was talking as the **active-AOI**. Each AOI

[1] https://www.fon.hum.uva.nl/praat/
[2] https://www.anvil-software.org/



```
for each participant:
{
    for each slide:
    {
        for each AOI:
        {
            1. Find start-time and end-time of the
               AOI from lecture annotation data.

            2. Extract all the fixations during that
               time, and calculate fixation duration.

            3. Find inside which AOI each fixation
               was located on the slide, and
               calculate the with-me-ness direction.

            4. Find the percentage of fixation
               duration in every direction, and
               create a direction sequence list.
        }
    }
}
```

Example for Pid1 for Slide Number 3, AOI=1

Step 1

Index	Start (seconds)	End (seconds)	transcript	AOI	Slide number
29	18.75137	19.10204	computer	1	3
30	19.10204	19.55408	data	1	3
31	19.88831	20.14309	is	1	3
...	...	...	...	...	...

AOI – 1
start_time = 18.75 sec
end_time = 29.17 sec

Step 2

f_no	f_x	f_y	f_start_time	f_end_time	fixation_dur
1	1420.475	578.9078	255	1059	804
2	740.66	570.7639	1059	1848	789
3	867.0255	130.3864	1848	2335	487
4	1286.209	438.7643	2335	3092	757
...	...	...	...	...	...
17	1230.022	611.0967	9239	9789	550

Step 3

f_no	f_x	f_y	f_start_time	f_end_time	fixation_dur	inside_aoi_1	inside_aoi_2	direction
1	1420.475	578.9078	255	1059	804	FALSE	TRUE	forward
2	740.66	570.7639	1059	1848	789	FALSE	TRUE	forward
3	867.0255	130.3864	1848	2335	487	FALSE	FALSE	out
4	1286.209	438.7643	2335	3092	757	TRUE	FALSE	same
...	...	...	...	...	...	...	...	
17	1230.022	611.0967	9239	9789	550	FALSE	TRUE	forward

Step 4

Pid	slide	active_aoi	time_same	time_forward	time_behind	time_outside	direction_sequence
Pid1	slide3	aoi1	0.364695	0.464233	NA	0.171072	forward-forward-outside-same------forward

Figure 3: The 4-step feature extraction algorithm used to extract active-AOI based gaze features for each active AOI.

on the slide was considered an active-AOI, as soon as the instructor started talking about it.

4.3.1 *With-me-ness directional features.*

To gain further insights into learners' viewing strategies, we propose a new set of features, which we call **"with-me-ness direction"**, which advances the concept of "with-me-ness" by not only explaining the amount of time learners' spent following the instructor's dialogues in a video-lecture but also in which direction they attend to the learning material. Based on the four possibilities of students' viewing behaviour, we defined four with-me-ness directions:

(1) **Same**: When the learners' gaze-points were present in the active-AOI (the AOI about which the instructor is talking). This also represents high Sharma et al. [22] conceptual with-me-ness (or high student-teacher co-attention)
(2) **Ahead**: When the learners' gaze-points were present in the AOI located *after* the active-AOI.
(3) **Behind**: When the learners' gaze-points were present in the AOI located *prior* to the active-AOI.
(4) **Outside**: When the learners' gaze-points were present *outside* of all pre-defined AOIs.

Further, to measure the directional aspect (i.e., in which direction the learner was looking when the instructor was talking about the active-AOI), and temporal aspect (i.e., how much time the learner spent looking in a direction when the instructor was talking about the active-AOI) of the with-me-ness direction, we defined two active-AOI based features:

- **With-me-ness direction sequences during Active-AOI**: This feature represents the transitions of learners' gaze direction while the instructor was talking about the active-AOI.
- **Fixation duration in every direction during Active-AOI**: This feature represents the amount of time spent in each direction, while the instructor was talking about the active-AOI.

4.3.2 *Feature-extraction algorithm.*

To extract the active-AOI based gaze features, we propose a four-step feature extraction algorithm. A graphical representation of the feature extraction algorithm is shown in Figure 3. As shown in the figure, for each participant, the following four steps were repeated per AOI per slide:

Step 1: Find start-time and end-time of the active AOI using lecture annotation data (Table 1).
Step 2: Extract all the fixations during that time, and calculate fixation duration.
Step 3: For each fixation, find which AOI it is located in, and in which direction. For instance, if active-AOI=1, and the fixation is located in AOI-2, then the direction of the gaze-vector is labelled as "ahead", as the participant is looking ahead than the active-AOI. Similarly, if active-AOI=1, and the fixation is located in AOI-1, then the direction is labelled as "same". Likewise, if active-AOI=2, and the fixation is located in AOI-1, then the direction is labelled as "behind". Lastly, if the fixation does not lie inside any of the predefined AOIs, then that fixation is labelled as "outside".
Step 4: Find the percentage of time a participant spent looking in each with-me-ness direction, and create a sequence list of the with-me-ness directions. In other words, we extracted the following five with-me-ness directional features:

– **time_same** : average time looking in the "same" direction.
– **time_ahead**: average time spent looking in the "ahead" direction.
– **time_behind**: average time spent looking in the "behind" direction.
– **time_outside**: average time spent in the "outside" direction.

- **direction_sequence**: sequence of with-me-ness directions

Figure 3 presents an example of how the algorithm extracts features. This example is for participant ID1, viewing slide 3, while the first AOI in the slide was active. The algorithm first extracts the time in the video lecture, where the instructor referred to the content of AOI-1. It then computes the start- and end-time based on the lecture annotation dataset (Table 1). The algorithm then extracts all the fixations that occurred during that time-span. The resulting dataset contains a list of fixations, their x-y coordinates, their start-time, and their end-time. We then calculate the duration of each fixation. Further, based on their x-y coordinates, we also detect in which AOI they are present, and compute their with-me-ness direction. In the final step, the algorithm extracts five *with-me-ness directional features* from the fixation dataset, i.e., calculate the fixation duration in every direction and direction sequence. The fixation duration is normalized by converting them from absolute values to the percentage of duration. For example, after Step 4, we find that 36.4% of the time they were looking in same AOI as active-AOI (time_same), 46.4% of the time they were looking ahead than active-AOI (time_ahead), and the remaining 17.2% of the time they were looking outside (time_outside). The direction sequences were created by combining all the gaze direction in a sequential manner.

The time-duration features (time_same, time_ahead, time_behind, time_outside) were averaged across all the slides, resulting in four feature values per participant. The sequence features (direction_sequence) were combined for each slide, resulting in 7 sequences (related to 7 slides) per participant.

4.4 Data Analysis Methods

To answer the first research question regarding inferring the video-watching behaviour from the sequences of learners' with-me-ness direction, we used *Discrete Time Markov Chains* (DTMC) for estimating the probability of the transitions between the with-me-ness directions. Further, to answer our second research question, we employed two techniques—(1) *Mediation analysis* for understanding the effect of prior knowledge on the relationship between learners' viewing strategies and their learning outcomes, and (2) *T-tests* to examine the difference in with-me-ness directions for novice and expert participants. We describe them briefly below.

4.4.1 Transition Matrix. Given a set of finite with-me-ness directions, the sequence of with-me-ness directions $X_1, X_2, X_3, ..., X_n$ is modeled as a 4-dimensional discrete time Markov Chain (DTMC) defined by the following states: $S = \{same, ahead, behind, outside\}$. The chain moves from one state to another, and the probability p_{ij} to move from state s_i to s_j in one step is known as the transition probability, and is defined as $p_{ij} = Pr(X_{t+1} = s_j | X_t = s_i)$. The matrix $P = (p_{ij})_{i,j}$, where each element of position (i, j) represents the transition probability p_{ij}, i.e., the transition matrix.

To understand learners' viewing strategies on a slide, the transition matrices for two groups of participants – high-achievers and low-achievers were compared first for a single slide and then for all the slides. The mean post-test score (*Mean* = 5.311, *SD* = 3.51) of the participants was used as the cut-off point to group them. We also performed t-test between the two groups, and report the result in terms of *p-value* statistic, *t-value* statistic and effect size (*Cohen's d*).

4.4.2 Mediation analysis. The mediation analysis proposed by Baron and Kenny [2] comprised of three sets of regression equations: IV → DV, IV → M, and IV + M → DV, where IV and DV are independent and dependent variables respectively, and M is the mediator variable. The first step is to conduct a simple regression analysis between the independent variable (IV) and the dependent variable (DV), to assess whether IV is a statistically significant predictor of DV ($Y = \beta_{10} + \beta_{11}X + \epsilon_2$). The criterion is satisfied if β_{11} is significant. The second step is to conduct a simple regression analysis between the IV and the mediating (M) variable, to assess whether IV is statistically significant predictor of M ($M = \beta_{20} + \beta_{21}X + \epsilon_2$). The criterion is satisfied if β_{21} is significant. Lastly, the third step is to conduct a multiple regression analysis with both the IV and the M in predicting the dependent variable (DV), to assess whether M is a significant predictor of Y, and whether the relationship between the IV and DV from Step 1 is significantly reduced or absent ($Y = \beta_{30} + \beta_{31}X + \beta_{32}M + \epsilon_3$). The criterion is satisfied if β_{32} is significant, and β_{31} should be less than the original relationship between IV and DV (β_{11}).

For mediation to occur, the three criteria described above should be fulfilled. Further, if the effect of IV on DV when M is added (β_{31} is non-significant) completely disappears, it is said that M fully mediates the relationship between IV and DV (full mediation). However, if the effect of IV on DV still exists, but in a smaller magnitude ($\beta_{31} < \beta_{32}$), M is said to partially mediate the relationship between IV and DV (partial mediation).

4.4.3 Testing differences between novice and expert groups. To test whether the with-me-ness directions were influenced by learners' prior knowledge, we compared the with-me-ness direction features between the novice group (participants with lower pre-test score) and expert group (participants with higher pre-test score). We used the mean pre-test score (*Mean* = 2.93, *SD* = 3.07) of all participants as the cut-off point. We performed t-test between the two groups, and report the result in terms of *p-value* statistic, *t-value* statistic and effect size (*Cohen's d*).

5 RESULTS

We examine the data at two levels for answering each research questions: (1) *slide-level* by analyzing learners' gaze transitions to answer RQ1, and (2) *lecture-level* by analyzing the impact of their overall gaze behaviour on their learning outcomes, and measuring the effect of prior knowledge on their with-me-ness direction features to answer RQ2.

5.1 Slide-Level Analysis: Understanding Learners' Gaze Transitions (RQ1)

Here, we describe how knowledge of learners' gaze transitions or gaze sequences can help us understand how they attend to the video lecture. For this, we compare the gaze sequences of the high (n = 24) and the low (n = 21) achieving students at a particular slide (slide 7) of the video lecture. This slide is divided into 3 AOIs, as shown in Figure 4.

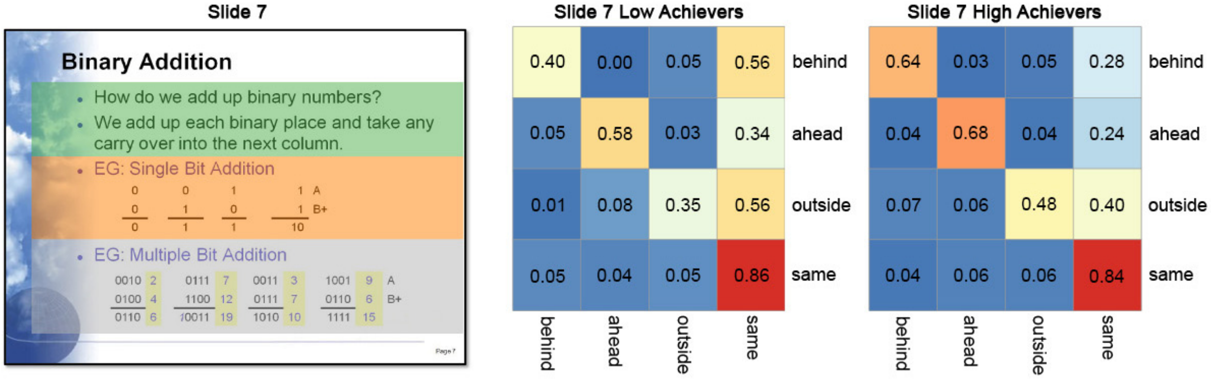


Figure 4: With-me-ness direction transitions for slide 7 for low achievers and high achievers. The label (behind, ahead, outside, and same) represent different directions. The numbers in the transition matrix defines the probability of a transition between row label (as source), and column label (as destination).

We found that both the low-achievers and the high-achievers were likely to follow the instructor and fixate within the active-AOI ($[same \longrightarrow same]_{low}$ = 86%, $[same \longrightarrow same]_{high}$ = 84%). Further, when the participants lagged behind, the probability for returning back to the active-AOI was higher for the low achievers ($[behind \longrightarrow same]_{low}$ = 56%) than for the high achievers ($[behind \longrightarrow same]_{high}$ = 28%). Similarly, when participants looked ahead, low-achievers were more likely to return to the active-AOI ($[ahead \longrightarrow same]_{low}$ = 34%), than the high-achievers $[ahead \longrightarrow same]_{high}$ = 24%). Likewise, when participants' gaze was outside of all defined AOIs, the probability of looking back to the active-AOI was higher for the low achievers ($[outside \longrightarrow same]_{low}$ = 56%), than that of the high achievers $[outside \longrightarrow same]_{high}$ = 40%). Lastly, the self-transitions for the high-achievers to stay looking in the ahead-AOI, behind-AOI or outside-AOI were higher than those of the low-achievers ($[ahead \longrightarrow ahead]_{low}$ = 58%, $[ahead \longrightarrow ahead]_{high}$ = 68%; $[behind \longrightarrow behind]_{low}$ = 40%, $[behind \longrightarrow behind]_{high}$ = 64%; $[outside \longrightarrow outside]_{low}$ = 35%, $[same \longrightarrow same]_{high}$ = 48%).

After analyzing learners' video watching behaviours at a particular slide, we wanted to investigate if these findings generalize for the complete video lecture. For this, we compared the high and the low achievers for all the seven slides in the lecture. Comparisons were made in terms of the probabilities of learners staying in a given AOI (self-transitions) or in terms of the probabilities of their returning back to active-AOI (i.e., looking in same-direction as the instructor's dialogues). The results are presented in Table 2.

The results shown in Table 2 are generally in line with the results shown in Figure 4, i.e., learners who were in an active-AOI tended to stay within the active-AOI (for both groups). Of the learners, who were in a 'behind' AOI, the high achievers were significantly more likely [T(df) = 2.48 (11.37), p-value = 0.03], to stay in the behind AOI, than the low achievers – who were significantly more likely [T(df) = -2.40 (9.25), p-value = 0.04] to transition from the 'behind' AOI to the active-AOI. For the remaining transitions, although the differences were not significant, generally, the low achievers tended to return to the active-AOI, and the high achievers tended to stay in the other AOIs.

5.2 Lecture-Level Analysis: Understanding the Relationship between Prior Knowledge, Learning Outcomes and With-me-ness Direction (RQ2)

After analyzing learners' "with-me-ness direction" sequences, we also analyzed whether the amount of time they spent looking in

Table 2: Statistical analysis results after comparing gaze transitions between low-achievers and high-achievers groups

	Low-Achievers (21)		*High-Achievers (24)*				
	M	*SD*	*M*	*SD*	*T (df)*	*p-value*	*Effect size*
$same \longrightarrow same$	0.881	0.077	0.866	0.061	-0.589 (11.994)	0.566	-0.315 (small)
$ahead \longrightarrow ahead$	0.573	0.071	0.636	0.074	1.625 (11.97)	0.130	0.869 (large)
$behind \longrightarrow behind$	0.508	0.078	0.601	0.061	2.482 (11.366)	0.030*	1.327 (large)
$outside \longrightarrow outside$	0.373	0.049	0.380	0.076	0.196 (10.227)	0.849	0.105 (negligible)
$behind \longrightarrow same$	0.434	0.092	0.339	0.050	-2.404 (9.249)	0.039*	-1.285 (large)
$ahead \longrightarrow same$	0.375	0.064	0.319	0.086	-1.358 (11.082)	0.201	-0.726 (medium)
$outside \longrightarrow same$	0.487	0.052	0.425	0.111	-1.342 (8.511)	0.214	-0.717 (medium)

$^{****}p < .0001$, $^{***}p < .001$, $^{**}p < .01$ and $^{*}p < .05$

Table 3: Descriptive statistics and correlation matrix

Features	*M*	*SD*	1	2	3	4	5	6
1. time_same	0.71	0.11	-					
2. time_ahead	0.16	0.08	-0.78****	-				
3. time_behind	0.08	0.05	-0.52***	-0.01	-			
4. time_outside	0.05	0.04	-0.51***	0.10	0.20	-		
5. prior knowledge	0.32	0.34	-0.45**	0.55****	0.10	-0.03	-	
6. learning outcomes	0.58	0.39	-0.39**	0.47**	0.12	-0.04	0.80****	-

****p < .0001, ***p < .001, **p < .01 and *p < .05

each direction throughout the video lecture (time_same, time_ahead, time_behind, time_outside) was related to their learning outcomes and prior knowledge.

5.2.1 Descriptive statistics and correlations. Table 3 shows descriptive statistics and correlations among the measures—prior knowledge, learning outcomes and four (eye-tracking related) 'with-me-ness direction' features. All variables are standardized. As shown in the table, on average, participants spent 71% of the time on the content instructor was talking about, 16% of time looking ahead from the instructor's dialogue, 8% looking behind than the instructor's dialogue and lastly, 5% time, looking outside the slide content or the pre-defined AOIs. The table further shows that the participants' learning outcomes were positively correlated with looking-ahead features (time_ahead), and negatively correlated with looking-same with-me-ness features (time_same). These correlations, however, should be interpreted with caution, as the relationship between these features with participants' learning outcomes could be influenced by their prior knowledge (r = 0.80, p<0.0001).

5.2.2 Mediation effect of prior knowledge on the relationship between with-me-ness direction and learning outcomes. To understand whether prior knowledge mediated the relationship between participants' learning outcomes and their looking ahead/same direction, we used Baron and Kenny [2] mediation analysis (described in Section 4.4.2), and defined the IV as looking ahead (time_ahead) and looking same (time_same), DV as learning outcomes (post-scores), and M as prior knowledge (pre-scores) respectively. The results from the mediation analysis in the form of three regression equations are present in Table 4. The table contains the values of the regression coefficients from the three regression equations. All variables shown are standardized.

The first regression equation analyzed the relationship between the with-me-ness direction and learning outcomes (IV → DV). It shows that a statistically significant relationship existed (β_{11} is significant, p<0.001). This satisfied the first criterion for mediation. The second regression equation (IV → M) suggests that learners with higher prior knowledge fixated more and spent more time on the contents presented ahead (r_{time_ahead} = 0.55, p< 0.0001), than the contents about which the instructor was talking (r_{time_same} = -0.45, p< 0.01). Prior knowledge having a positive significant association with the 'ahead' with-me-ness feature (β_{21}>0 and significant, p< 0.0001) and a significant negative association with the 'same' with-me-ness feature (β_{21}<0 and significant, p< 0.001) satisfied the second criterion for mediation. Finally, the results from the third regression equation suggest that prior knowledge **"fully mediated"** the relationship between the with-me-ness features and learning outcomes (β_{32} was significant, p<0.0001 and β_{31} was non-significant). This implies that there is no direct-effect of participants' viewing behaviour on their learning outcomes, the observed change in viewing behaviour (such as looking ahead, looking same) was largely influenced by their prior knowledge.

5.2.3 The relationship between with-me-ness direction and prior knowledge. This section reports the results of the analysis that evaluated *how* learners with different levels of prior knowledge watched the video lecture, and *what* strategies did they follow to achieve better learning outcomes. Table 5 shows the result after comparing the with-me-ness direction between novice (n = 23) and expert (n = 22) groups.

We found that there was a significant difference between these groups in terms of how they watched the video lecture. As compared to the novice learners, expert participants spent lesser time in the

Table 4: Regression coefficients after the mediation analysis among with-me-ness direction, prior knowledge and learning outcomes

Independent Variable (IV)	*Step 1: IV->DV* (β_{11})	*Step 2: IV->M* (β_{21})	*Step 3: (IV+M->DV)*		*Mediation analysis result*
			IV (β_{31})	*M* (β_{32})	
Percentage of time spent in ahead AOIs (t_ahead)	2.188***	2.268****	0.181	0.885****	Full mediation
Percentage of fixations in same AOIs (f_same)	-1.384**	-1.394***	-0.147	0.888****	Full mediation

****p < .0001, ***p < .001, **p < .01 and *p < .05

Table 5: Statistical analysis results after comparing novice and expert students groups in terms of their with-me-ness direction.

	Novice Students (23)		*Expert Students (22)*				
	M	*SD*	*M*	*SD*	*T (df)*	*p-value*	*Effect size*
time_same	0.748	0.073	0.663	0.127	2.750 (33.14)	0.0095**	0.830 (large)
time_ahead	0.117	0.056	0.212	0.080	-4.587 (37.62)	4.8e-5****	-1.378 (large)
time_behind	0.077	0.048	0.075	0.055	0.148 (33.86)	0.882	0.004 (negligible)
time_outside	0.058	0.027	0.050	0.046	0.635 (38.88)	0.529	0.191 (negligible)

****p < .0001, ***p < .001, **p < .01 and *p < .05

same AOIs– about which the instructor was talking ($\text{time\_same}_{novice} = 74.8\%$, $\text{time\_same}_{expert} = 66.3\%$, $t(33.14) = 2.750, p < 0.01$). Moreover, expert learners also spent significantly more time in the ahead-AOI than the novice learners ($\text{time\_ahead}_{novice} = 11.7\%$, $\text{time\_ahead}_{expert} = 21.2\%$, $t(37.62) = -4.587, p < 0.0001$).

6 DISCUSSION

Our analysis sheds light on learners' video viewing strategies using our proposed metric –"with-me-ness direction". We found that participants mostly followed the instructor's dialogue ($time\_same$ = 71%), and their gaze mostly fixated on the active-AOI (the same area on the video-lecture slide about which the instructor was talking). The feature 'time_same' reflects Sharma et al. [22]'s conceptual "with-me-ness", which suggests that the participants attentively watched the video-lecture during the study. However, measuring the direction of learners' gaze using our proposed metric helped us to build a more comprehensive picture of learners' visual attention allocation during video-watching.

RQ1: What can we infer about learners' video-watching behaviour from the sequences of 'with-me-ness direction'?

Our results highlight a range of different visual behaviours while the learners watched a video lecture. Although all participants were more likely to fixate within the active-AOI than anywhere else, the participants exhibited nuanced differences in their behaviours (e.g. by looking away from the screen while listening to the lecture, looking ahead, or revisiting AOIs that had already been discussed). High- and low-achievers, as determined by their post-test scores, exhibited different visual behaviours. First, the low-achievers spent more time following the instructor's dialogues. Second, if the low achieving learners' attention was directed to other areas on the slide, the tendency of them looking again in the active-AOI and paying attention to instructor's dialogues was more than the high achievers. This behaviour could be indicative that these learners were trying to understand the learning material (as they had lower prior knowledge). Third, the high-achievers tended to explore the slide content more by looking ahead, looking outside or reviewing the content that had already been discussed. Their tendency of looking more in the behind-direction than the low-achievers, is contrary to our expectations. Due to their higher prior knowledge, we hypothesized that high-achievers would be able to build mental models of the presented information more easily and thus, may require fewer transitions in the behind-AOI. However, their behaviour of fixating more on non-relevant areas, and showing lesser gaze-dispersion (low tendency to transit back to active-AOI from behind-AOI) could be reflective of mind-wandering state [31]. In future, to better understand these behaviours, learners could be encouraged to self-report their reasons for looking 'outside', 'ahead' or 'behind' via video-simulated recall.

RQ2: Is there a relationship between learners' with-me-ness direction, their learning outcomes and prior knowledge?

Sharma et al. [22] in their study reported that students' "with-me-ness" is positively correlated with their learning outcomes. In our study, we found that prior knowledge can fully mediate the relationship between participants' learning outcomes and their gaze with-me-ness direction (see Table 4). A possible explanation for this could be that participants' prior knowledge was a strong positive predictor of their learning outcomes. Therefore, the participants who had some prior knowledge of binary numbers could answer the pre-test questions without watching the video lecture. Further, we found that prior knowledge had a significant impact on participants' video-watching behaviours. The participants with higher prior knowledge were looking more on the content presented *ahead* in the slides than on the content about which the instructor was talking. However, the participants with lower prior knowledge seemed to have followed the instructor's dialogue (see Table 5 for details). These findings are consistent with previous research which shows that learners with higher prior knowledge attend more to relevant information (see Section 2.2.1), which might have resulted in expert learners' looking more in the forward-direction. However, as acknowledged by Sinha et al. [23], the cognitive resource allocation during video-watching also depends on two additional factors – (1) perceived difficulty, and (2) motivation. Sharma et al. [21] have shown that motivation can effect learners' with-me-ness (i.e., looking in same-direction), however, to confirm the reason that the learners were looking in the ahead-direction was due to their high-prior knowledge or content being too easy or that they lost the motivation, further research is required.

One of our aims in this paper is to demonstrate the usefulness of eye-tracking measures in achieving new insights about learners' attention patterns. However, there were a few inherent limitations arising from the dataset employed. First, the participants were not permitted to rewind, pause or replay the videos and were continuously self-reporting their perceived difficulty using a slider while watching the video-lecture. Srivastava et al. [26] suggests that this type of data collection should not incur additional cognitive load, we do consider the possibility that self-reporting might have influenced their video-watching behaviours. Second, we realize there is a tradeoff between the scalability and reliability of our methods, therefore, in future work, we plan to extend this study and test this metric across a range of slideware/video-lecture designs.

7 CONCLUSION

In this paper, we have proposed a new gaze-based metric "with-me-ness direction", that can measure the co-attention between learners' gaze and instructor's dialogues (same-direction), and can also compute the attentional direction of learners (ahead, behind and outside) during the instructor's dialogues in a video-lecture. We found that most of the time learners fixated on the same area of the video-lecture slide which the instructor was talking about. However, the tendency of fixating more on the new content of the slide, and showing lesser transitions back to the active-AOI (the area of the slide about which the instructor was talking) was higher for learners with higher prior knowledge than those with lower prior knowledge. To sum up, we have found that the co-attention direction is useful in understanding how learners may interact with video-lectures, and it provides an informative window into their video-watching behaviours. These findings, together with our proposed methodology present an opportunity to utilize eye-tracking based techniques for preparing and redesigning high-production educational videos, as well as to compare different slideware styles that could assist in the design of optimal instructional material. Therefore, this paper has direct implications for instructors for the design of effective learning materials.

REFERENCES

[1] Ecenaz Alemdag and Kursat Cagiltay. 2018. A systematic review of eye tracking research on multimedia learning. *Computers & Education* 125 (2018), 413 – 428. https://doi.org/10.1016/j.compedu.2018.06.023

[2] Reuben M. Baron and David A. Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51, 6 (1986), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

[3] Matt Canham and Mary Hegarty. 2010. Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction* 20, 2 (2010), 155 – 166. https://doi.org/10.1016/j.learninstruc.2009.02.014

[4] Michail N. Giannakos, Konstantinos Chorianopoulos, and Nikos Chrisochoides. 2015. Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. *The International Review of Research in Open and Distributed Learning* 16, 1 (Jan. 2015). https://doi.org/10.19173/irrodl.v16i1.1976

[5] Michail N. Giannakos, Demetrios G. Sampson, and Łukasz Kidziński. 2016. Introduction to smart learning analytics: foundations and developments in video-based learning. *Smart Learning Environments* 3, 1 (07 Jul 2016), 12. https://doi.org/10.1186/s40561-016-0034-2

[6] Stephen Hutt, Jessica Hardey, Robert Bixler, Angela Stewart, Evan Risko, and Sidney K D'Mello. 2017. Gaze-Based Detection of Mind Wandering during Lecture Viewing. *International Educational Data Mining Society* (2017).

[7] DongMin Jang, IlHo Yang, and SeoungUn Kim. 2020. Detecting Mind-Wandering from Eye Movement and Oculomotor Data during Learning Video Lecture. *Education Sciences* 10, 3 (2020), 51. https://doi.org/10.3390/educsci10030051

[8] Halszka Jarodzka, Katharina Scheiter, Peter Gerjets, and Tamara van Gog. 2010. In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction* 20, 2 (2010), 146 – 154. https://doi.org/10.1016/j.learninstruc.2009.02.019

[9] Marcel A. Just and Patricia A. Carpenter. 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review* 87, 4 (1980), 329.

[10] Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller. 2003. The Expertise Reversal Effect. *Educational Psychologist* 38, 1 (2003), 23–31. https://doi.org/10.1207/S15326985EP3801_4

[11] Juho Kim, Philip J. Guo, Daniel T. Seaton, Piotr Mitros, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Understanding In-Video Dropouts and Interaction Peaks Inonline Lecture Videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference* (Atlanta, Georgia, USA) *(L@S '14)*. Association for Computing Machinery, New York, NY, USA, 31–40. https://doi.org/10.1145/2556325.2566237

[12] René F. Kizilcec, Kathryn Papadopoulos, and Lalida Sritanyaratana. 2014. Showing Face in Video Instruction: Effects on Information Retention, Visual Attention, and Affect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2095–2102. https://doi.org/10.1145/2556288.2557207

[13] Joy Yeonjoo Lee, Jeroen Donkers, Halszka Jarodzka, and Jeroen J.G. van Merriënboer. 2019. How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking. *Computers in Human Behavior* 99 (2019), 268 – 277.

[14] Jason Lodge, Jared Cooney Horvath, Alex Horton, Gregor Kennedy, Sven Venema, and Shane Dawson. 2017. Designing videos for learning: Separating the good from the bad and the ugly. (01 2017).

[15] R. Mayer and R.E. Mayer. 2005. *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press.

[16] Richard E. Mayer. 2002. Multimedia learning. , 85 - 139 pages. https://doi.org/10.1016/S0079-7421(02)80005-6

[17] Pontus Olsson. 2007. Real-time and Offline Filters for Eye Tracking.

[18] Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology* 45, 3 (1991), 255.

[19] Marija Sablić, Ana Mirosavljević, and Alma Škugor. 2020. Video-Based Learning (VBL)—Past, Present and Future: an Overview of the Research Published from 2008 to 2019. *Technology, Knowledge and Learning* (07 Jul 2020). https://doi.org/10.1007/s10758-020-09455-5

[20] Amir Shareghi Najar, Antonija Mitrovic, and Kourosh Neshatian. 2015. Eye tracking and studying examples: how novices and advanced learners study SQL examples. *Journal of computing and information technology* 23, 2 (2015), 171–190. https://doi.org/10.2498/cit.1002627

[21] Kshitij Sharma, Michail Giannakos, and Pierre Dillenbourg. 2020. Eye-tracking and artificial intelligence to enhance motivation and learning. *Smart Learning Environments* 7, 13 (2020), 1–19. https://doi.org/10.1186/s40561-020-00122-x

[22] Kshitij Sharma, Patrick Jermann, and Pierre Dillenbourg. 2014. "With-me-ness": A gaze-measure for students' attention in MOOCs. In *Proceedings of International Conference of the Learning Sciences 2014*. ISLS, 1017–1022.

[23] Tanmay Sinha, Patrick Jermann, Nan Li, and Pierre Dillenbourg. 2014. Your click decides your fate: Leveraging clickstream patterns in MOOC videos to infer students' information processing and attrition behavior. *CoRR* abs/1407.7131 (2014). arXiv:1407.7131 http://arxiv.org/abs/1407.7131

[24] Namrata Srivastava, Sadia Nawaz, Jason M. Lodge, Eduardo Velloso, Sarah Erfani, and James Bailey. 2020. Exploring the Usage of Thermal Imaging for Understanding Video Lecture Designs and Students' Experiences. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (Frankfurt, Germany) *(LAK '20)*. ACM, New York, NY, USA, 250–259. https://doi.org/10.1145/3375462.3375514

[25] Namrata Srivastava, Joshua Newn, and Eduardo Velloso. 2018. Combining Low and Mid-Level Gaze Features for Desktop Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 189 (Dec. 2018), 27 pages. https://doi.org/10.1145/3287067

[26] Namrata Srivastava, Eduardo Velloso, Jason M. Lodge, Sarah Erfani, and James Bailey. 2019. Continuous Evaluation of Video Lectures from Real-Time Difficulty Self-Report. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300816

[27] Michelle Taub and Roger Azevedo. 2016. Using Eye-Tracking to Determine the Impact of Prior Knowledge on Self-Regulated Learning with an Adaptive Hypermedia-Learning Environment. In *Proceedings of the 13th International Conference on Intelligent Tutoring Systems - Volume 9684 (ITS 2016)*. Springer-Verlag, Berlin, Heidelberg, 34–47. https://doi.org/10.1007/978-3-319-39583-8_4

[28] Tamara van Gog and Halszka Jarodzka. 2013. *Eye Tracking as a Tool to Study and Enhance Cognitive and Metacognitive Processes in Computer-Based Learning Environments*. Springer New York, New York, NY, 143–156. https://doi.org/10.1007/978-1-4419-5546-3_10

[29] Jeffrey D. Wammes, Paul Seli, J. Allan Cheyne, Pierre O. Boucher, and Daniel Smilek. 2016. Mind wandering during lectures II: Relation to academic performance. *Scholarship of Teaching and Learning in Psychology* 2, 1 (2016), 33–48. https://doi.org/10.1037/stl0000055

[30] Jiahui Wang and Pavlo D. Antonenko. 2017. Instructor presence in instructional video: Effects on visual attention, recall, and perceived learning. *Computers in Human Behavior* 71 (2017), 79 – 89. https://doi.org/10.1016/j.chb.2017.01.049

[31] Han Zhang, Kevin F. Miller, Xin Sun, and Kai S. Cortina. 2020. Wandering eyes: Eye movements during mind wandering in video lectures. *Applied Cognitive Psychology* 34, 2 (2020), 449–464. https://doi.org/10.1002/acp.3632