The Effects of Video Instructor’s Body Language on Students’ Distribution of Visual Attention: an Eye-tracking Study

Jiawen Zhang
Beijing University of Posts and Telecommunications
zhangjiawen@bupt.edu.cn

Marie-Luce Bourguet
Queen Mary University of London
marie-luce.bourguet@qmul.ac.uk

Gentiane Venture
Tokyo University of Agriculture and Technology
venture@cc.tuat.ac.jp

Previous studies have shown that the instructor’s presence in video lectures has a positive effect on learners’ experience. However, it does increase the cost of video production and may increase learners’ cognitive load. An alternative to instructor’s presence is the use of embodied pedagogical agents that display limited but appropriate social signals. In this extended abstract, we report a small experimental study into the effects of video instructor’s behaviour on students’ learning experience, with the long term aim of better understanding which instructor’s social signals should be applied to pedagogical agents. We used eye-tracking technology and data visualisation techniques to collect and analyse students’ distribution of visual attention in relation to the instructor’s speech and body language. Participants also answered questions about their attitudes toward the instructor. The results suggest that the instructor’s gaze directed towards the lecture’s slides, or a pointing gesture towards the slides, is not enough to shift viewers’ attention. However, the combination of both is effective. An embodied pedagogical agent should be able to display a multimodal behaviour, combining gaze and gestures, to effectively direct the learners’ visual attention towards the relevant material. Furthermore, to make learners pay attention to the lecturer’s speech, the instructional agent should make use of pauses and emphasis.

Video lectures. Social signals. Eye tracking. Embodied pedagogical agents.

1. INTRODUCTION

In remote learning, videos have the potential to offer many of the advantages of a classroom-like experience and, in addition, they enable student’s control over the pace of their learning (Yousef et al., 2014). Various studies have looked at the effects of different video-based instruction designs in relation to students’ engagement, attention, emotion, cognitive load, knowledge transfer and recall (Chen & Wu, 2015; Guo et al., 2014). Based on the eye-mind assumption that eye fixation locations reflect attention distributions (Just & Carpenter, 1980), an increasing number of studies are using eye-tracking techniques to understand how students learn using videos (Lai et al., 2013; Sharma et al., 2014) and especially how instructor’s presence in the video affects students distribution of visual attention (Garrett, 2015; Kizilcec et al., 2014).

A general positive effect of instructor’s presence in instructional videos has been found (Wang & Antonenko, 2017). For example, it contributes to increase students’ “with-me-ness”, which is the extent to which the learner succeeds in following the content that is being explained (Sharma et al., 2016). Moreover, as lecturers’ hand gestures and facial expressions are often linked to their pedagogical intentions (Tian & Bourguet, 2016; Zhang, 2012), the availability of social signals such as the instructor’s pointing gestures and gaze can improve learning experience and performance (Ouwehand et al., 2015; Pi et al., 2017).

However, including the lecturer’s presence in videos entails a high production cost (Hollands & Tirthali, 2014). Moreover, there is a concern that it may contribute to increasing the learners’ cognitive load (Chandler & Sweller, 1991; Mayer, 2001) by inducing a split attention effect (when learners must divide their attention across multiple information sources). For example, it has been found that learners are looking at the instructor’s face up to 65% of the time in average, and that they switch between the lecturer’s face and the instructional material up to every 2.4 seconds, depending on the multimedia design (Garett, 2015).

A low-cost and accessible alternative to instructor’s presence in videos is the use of embodied pedagogical agents (Li et al., 2015). Agents that display limited but appropriate social signals may also incur less cognitive load than their human

Proceedings of British HCI 2018. Belfast, UK.

models. In this work-in-progress paper, we report the results of an experimental study into the effects of video instructor’s behaviour on students’ learning experience, with the long term aim of better understanding which instructor’s social signals should be applied to pedagogical agents. The scale of the study is small (8 participants), but at this early stage of the research, the intention is to capture some of the instructor’s important social signals in order to build a first prototype that can be used for further studies. We briefly describe our pedagogical agent prototype in the conclusion of the paper.

2. EYE-TRACKING EXPERIMENT

2.1 Method

We used eye tracking technology and data visualisation techniques (Bojko, 2009) to collect and analyse students’ distribution of visual attention in relation to the video instructor’s speech and body language. Participants to the experiment also answered questions about their attitudes toward the instructor.

2.1.1. Video Stimulus
All participants watched the same video (duration of 4 minutes and 13 seconds) on the topic of “Design Techniques” (covering brain storming, mind maps and storyboards), extracted from a 3rd year undergraduate telecommunications engineering course. The video showed the instructor’s head and upper body on the right side of the lecture’s slides, all within the same frame (see Figure 1).

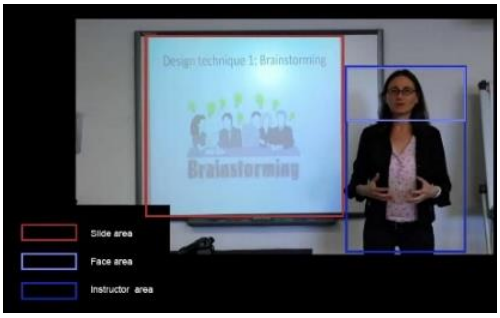

Figure 1: *The Areas of Interest (AOI).*

Prior to conducting the experiment, the video was manually annotated with instructor behaviour’s markers, using the ANVIL annotation tool (Kipp, 2014). Behaviour markers included three markers for gaze (looking towards the camera, i.e. the viewer; looking towards the slides; looking elsewhere); seven markers for hand gestures (pointing towards slide; waving hands; clasping hands; unfolding hands; ball; other gesture; no gesture); and three markers for speech (speaking with direct reference to the slide’s content; not directly referring to slide’s content; no speech). The annotations were not displayed to the participants.

2.1.2. Participants
Ten undergraduate students from an International Bachelor’s degree in Electronic Engineering in China delivered in the English language were recruited. Prior to the study, each participant was asked to complete a background questionnaire to ensure that all participants shared a similar level of prior domain knowledge (all of them had taken the module of the video in the previous semester) and English comprehension (CET6 level). Two participants had to be excluded due to problems with their eye tracking data, leaving a sample of eight participants, three males and five females, aged 20 to 22. None of them had abnormal vision or abnormal hearing.

2.1.3. Procedure and Equipment
The experiment was conducted in individual sessions of approximately 10 minutes. Before the video stimulus started, the experimenter gave participants a brief introduction to the experiment and to the eye-tracking equipment, and each participant was asked to follow a simple procedure for equipment calibration purpose. The participants were then asked to watch the instructional video without being able to pause or stop it. To ensure that they were paying attention and trying to learn from the video, they were told that they would have to write a summary of the video content immediately after watching it.

The participants’ eye position was measured using the Tobii 4C eye tracker. The device was mounted to the bottom of the computer monitor on which the lecture video was displayed. Tobii 4C operates at a distance of 50-95cm and has a high accuracy of 0.4 degrees. The sampling frequency is 90 Hz. Computer’s screen size was 13.3 inches, and the resolution of the monitor was 1440 x 900 pixels.

2.2 Measurements

Visual attention is typically measured in the form of fixations, which (in our study) describe durations of at least 200ms that a viewer spends looking at a small area on the screen (i.e. an area of side limited to 10 pixels). Fixations are connected by saccades, and a sequence of fixations and saccades is called a scanpath.

2.2.1. Areas of Interest
Areas of Interest (AOIs) are parts of the video frame that are of high importance for the hypothesis of the study. Two non-overlapping AOIs were determined: the instructor area and the slide area (see Figure 1).

We found that, in average, participants spend 95.33% of their time (percentage of gaze point distribution) watching one of the two AOIs. They spend slightly more time on the instructor AOI (M=49.09%, SD=14.12) than on the slide AOI (M=46.23%, SD=13.58), although the difference is not significant (a paired sample t-test was conducted: t(7) = 0.29, p=0.05, ns). After dividing the instructor AOI into two: the face area and the body area; we observed that students look more at the instructor's face than the body gestures (M=75.66%, SD=11.28).

Table 1 shows for each AOI and different instructor behaviours the average fixation rate, i.e. the average fixations count divided by the total duration of the behaviour (note that the gaze, hand and speech behaviours are not exclusive behaviours). Surprisingly, behaviours that are meant to attract attention to the lecture's slide (e.g. Gaze towards slide, Hand pointing and Speech with reference) have higher fixation rate on the instructor AOI than on the slide AOI. This could be explained by the fact that there is a delay between the behaviour and the effect it has on the students' visual attention. It could also be explained by a higher rate of transitions between the two AOIs (see next section). For a better explanation, combinations of behaviours should in fact be scrutinised (see visualisation section).

***Table 1:** Average fixation rate (count / duration) on each AOI in relation to instructor behaviour; and total duration (in seconds) of the observed behaviours.*

Behaviour	Instruct. AOI	Slide AOI	Duration (seconds)
Gaze towards camera	*0.413*	*0.607*	*226.8*
Gaze towards slide	*0.62*	*0.219*	*10.88*
Gaze other	*0.681*	*0.157*	*14.32*
Hand pointing	*0.560*	*0.351*	*26.36*
Hand waving	*0.338*	*0.593*	*70.60*
Unfolding hands	*0.248*	*0.654*	*31.72*
Other hand gesture	*0.247*	*0.390*	*99.32*
No hand gesture	*0.354*	*0.245*	*24.00*
Speech with reference	*0.508*	*0.437*	*58.04*
Speech no reference	*0.369*	*0.623*	*166.92*
Silence	*0.579*	*0.339*	*30.24*

2.2.2. Transitions

A transition is a movement from one AOI to another. The typical measure related to transitions is the transition count, i.e. number of transitions between two AOIs.

Table 2 shows average transition counts across participants in relation to different instructor's behaviours. Given that the total duration of each behaviour is variable, we computed an average transition rate (transition count/duration) for each behaviour. We can see that when the instructor is looking at the slide, the transition rate is relatively high (1.195), which contributes to shorten the fixation rate on the slide AOI and corroborates the findings of Table 1.

Table 2: Average transition count [standard deviation] and transition rate (count/duration) between the two AOIs in relation to instructor behaviour.

Behaviour	Average transition count [SD]	Average transition rate
Gaze towards camera	*207.38 [70.96]*	*0.914*
Gaze towards slide	*13.00 [7.05]*	*1.195*
Gaze other	*8.25 [4.29]*	*0.576*
Hand pointing	*24.88 [10.15]*	*0.944*
Hand waving	*64.63 [21.81]*	*0.915*
Unfolding hands	*32.25 [11.79]*	*1.017*
Other hand gesture	*78.38 [25.63]*	*0.789*
No hand gesture	*24.63 [13.35]*	*1.026*
Speech with reference	*51.63 [17.07]*	*0.890*
Speech no reference	*152.88 [54.39]*	*0.916*
Silence	*23.25 [9.48]*	*0.769*

2.3 Visualisation

2.3.1 Attention maps

An attention map (or heat map) is a graphical representation of the attention distribution. Different kinds of attention maps have been proposed (Bojko, 2009), e.g.: "Fixation count heat map", which results from the aggregation of fixation counts across time and participants (also called bee swarm); and "Absolute gaze duration heat map", which is the aggregation of absolute gaze duration across time and participants.

Figure 2 (left) shows a fixation count heat map calculated on a 5.32 second clip during which the instructor is performing a pointing gesture, looking at the slide and delivering speech that is directly referring to the slide content. With the three behaviours combined, the student's visual attention is clearly directed towards the slide AOI, where gaze duration is also longer.

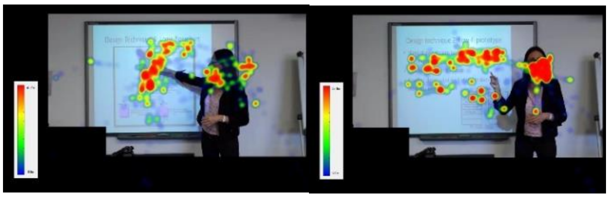

***Figure 2:** Fixation count heat maps where the instructor looks at the slide (left) versus the camera (right).*

Figure 2 (right) shows a fixation count heat map calculated on a 4.52 second clip during which the instructor is performing a pointing gesture and delivering speech that is directly referring to the slide content but is looking at the camera. The student's visual attention is scattered on the slide, as if the gesture alone did not allow them to find the relevant information, and the gaze duration is actually longer on the instructor's face.

2.3.2 Temporal Evolution of Scanpaths

Figure 3 shows horizontal fixation positions in the vertical axis and time on the horizontal axis. Each line corresponds to a different participant. The top image has been calculated on the clip of Figure 2 (left) (slightly extended to 6 second duration), during which the instructor is pointing at the slide while looking towards it. The bottom image has been calculated on the clip of Figure 2 (right) (also extended to the same 6.00 second duration), during which the instructor is pointing at the slide and looking at the audience.

We can clearly observe less transitions between the two AOIs in the top image, showing that the instructor's gaze towards the slide, when combined with a pointing gesture, has the effect of helping students maintain their attention on the slide. The pointing gesture alone does not prevent students from shifting their attention back and forth between the slide and the instructor's face, hence potentially increase their cognitive load. In the bottom image, some students keep staring at the instructor's face, whereas in the top image the opposite can be observed: some students keep staring at the slide without shifting their attention back to the instructor.

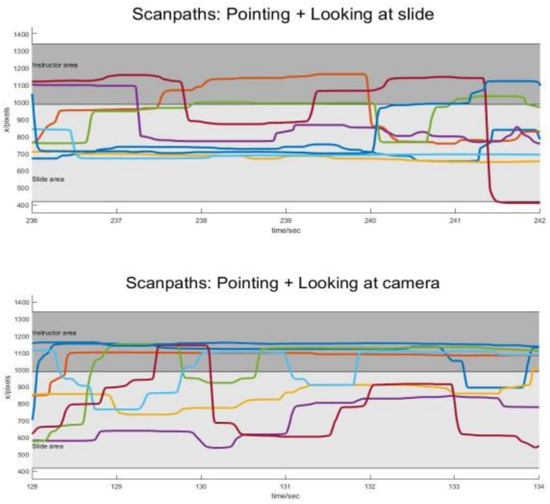

Figure 3: *Scanpaths when the instructor is looking at the slide (top image) versus the camera (bottom image). The instructor AOI is the top dark grey area.*

3. QUESTIONNAIRE RESULTS

After watching the video, the participants answered a short questionnaire about their attitude towards the video instructor.

The results show that all participants consider the instructor's presence useful, and 75% of them think that the instructor's behaviour is helping them understand the lecture's content. Two behaviours in particular: 'Hand pointing' and 'Speech with emphasis', are regarded as particularly important. When the instructor performs a pointing gesture, 87.5% of the participants thought that there must be something worth of attention in the slide, which corroborates the results of the eye tracking experiment. Conversely, 62.5% of the participants believed that 'Speech with emphasis' means that the instructor was saying something important.

Further results show that participants feel most concerned with the instructor's speech, followed by the slides' area, and finally the instructor's body. Indeed, we know already from the eye tracking experiment that participants spend much more time looking at the instructor's face area than the body area. The main function of the instructor's behaviour is to help shifting the students' visual attention between the teaching material and the teacher's face, i.e. the speech.

4. CONCLUSION AND FURTHER WORK

In this paper, we reported a small experimental study into the effects of video instructor's behaviour on students' distribution of visual attention. The results suggest that pointing gestures combined with gaze constitute an important and useful social signal. An embodied pedagogical agent should be able to display a multimodal behaviour, combining gaze and gestures, to effectively direct the learners' visual attention towards the relevant material. Furthermore, to make learners pay attention to the speech, the instructional agent should make use of pauses and emphasis.

We have implemented a prototype of an embodied pedagogical agent for further studies on what social signals should such an agent display (Figure 4). We chose the social robot Pepper (SoftBank Robotics, 2017) because of its neutrality (e.g. it is non-gendered), because a robot looks playful and non-judgmental (Clark & Mayer, 2011), and because it is not expected to display the complex, but not always useful, behaviour of a human instructor. Pepper's main social signals for now include gaze (head direction) and pointing gestures. Further studies using Pepper are being conducted to test the acceptability of a robot as instructor, and the social signals it should display to support the learners.

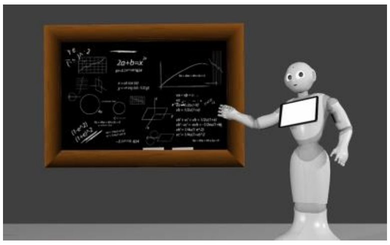

Figure 4: *Pepper the virtual social robot and embodied pedagogical agent.*

5. REFERENCES

Bojko, A. (2009). Informative or Misleading? Heatmaps Deconstructed. Human-Computer Interaction. New Trends. Springer Berlin Heidelberg.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition & Instruction, 8*(4), 293-332.

Chen, C. M., & Wu, C. H. (2015). Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. *Computers & Education, 80*(5), 108-121.

Clark, R. C., & Mayer, R. E. (2011). *E-learning and the science of instruction: proven guidelines for consumers and designers of multimedia learning*. Pfeiffer.

Garrett, N. (2015). Eye-Tracking Analytics in Instructional Videos. *ISECON*.

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: an empirical study of MOOC videos. *ACM Conference on Learning @ Scale Conference* (Vol.43, pp.41-50). ACM.

Hollands, F.M., & Tirthali, D. (2014). MOOCs: Expectation sand reality. Full Report. Center for Benefit Cost Studies of Education, Teachers College, Columbia University, NY.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review, 87*(4), 329.

Kipp, M. (2014). ANVIL: A Universal Video Research Tool. In J. Durand, U. Gut, G. Kristofferson (Eds.) *Handbook of Corpus Phonology*, Oxford University Press, 420-436.

Kizilcec, R. F., Papadopoulos, K., & Sritanyaratana, L. (2014). Showing face in video instruction: effects on information retention, visual attention, and affect. 2095-2102.

Lai, M. L., Tsai, M. J., Yang, F. Y., Hsu, C. Y., Liu, T. C., & Lee, W. Y., et al. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review, 10*(4), 90-115.

Li, J., Kizilcec, R., Bailenson, J., & Ju, W. (2015). Social robots and virtual agents as lecturers for virdo instruction. *Computers in Human Behavior*, *55*, 1222-1230.

Mayer, R. E. (2001). *Multimedia Learning*. Cambridge University Press.

Ouwehand, K., Van Gog, T., & Paas, F. (2015). Designing effective video-based modeling examples using gaze and gesture cues. *Educational Technology & Society, 18*.

Pi, Z., Hong, J., & Yang, J. (2017). Effects of the instructor's pointing gestures on learning performance in video lectures. *British Journal of Educational Technology, 48*(4), 1020-1029.

Sharma, K., Jermann, P., & Dillenbourg, P. (2014). How Students Learn using MOOCs: An Eye-tracking Insight. *EMOOCs 2014, the Second MOOC European Stakeholders Summit*.

Sharma, K., Alavi, H. S., Jermann, P., & Dillenbourg, P. (2016). A gaze-based learning analytics model:in-video visual feedback to improve learner's attention in moocs. 417-421.

SoftBank Robotics (2017) "Find out more about Pepper". [Online] Available from: https://www.ald.softbankrobotics.com/en/robots/pepper [accessed 28 March 2018].

Tian, Y., & Bourguet, M. L. (2016). Lecturers' Hand Gestures as Clues to Detect Pedagogical Significance in Video Lectures. *European Conference on Cognitive Ergonomics* (pp.2). ACM.

Wang, J., & Antonenko, P. D. (2017). Instructor presence in instructional video: effects on visual attention, recall, and perceived learning. *Computers in Human Behavior, 71*, 79-89.

Yousef, A. M. F., Chatti, M. A., & Schroeder, U. (2014). Video-Based Learning: A Critical Analysis of The Research Published in 2003-2013 and Future Visions. *eLmL 2014 : The Sixth International Conference on Mobile, Hybrid, and On-line Learning* (pp.112-119).

Zhang, J. R. (2012). Upper body gestures in lecture videos:indexing and correlating to pedagogical significance. *ACM International Conference on Multimedia* (pp.1389-1392). ACM.