

What Matters most for Learning from Online Videos, Seeing the Instructor's Face or Gaze? Impact on Instructor-Student Synchrony

Bertrand Schneider, Harvard University, bertrand_schneider@gse.harvard.edu

Gahyun Sung, Harvard University, gsung@g.harvard.edu

Javaria Hassan, Harvard University, javariahassan@gse.harvard.edu

Harum Bhinder, Harvard University, harum_bhinder@gse.harvard.edu

Abstract: Over the last decade, the prevalence of online teaching has dramatically increased. As part of their learning experience, students are expected to spend more and more time watching online videos. In this paper, we explore alternative formats for designing these videos. We use theory-driven learning principles to create videos that can specifically support students' learning (i.e., by leveraging social theories of learning showing that joint visual attention plays a central role in knowledge building). We compare the benefits of seeing the instructor's face and / or gaze on instructional videos using a 2x2 experimental design. We found that showing the instructor's face had no significant effect on learning, while adding the instructor's eye-tracking data to the video promoted conceptual understanding. We tested several mediators for learning, such as joint visual attention, joint emotional response, and movement synchrony; we found that joint visual attention played a significant mediatory role. We conclude by discussing the implications of these findings and recommendations for designing online learning videos.

Introduction

Online learning has shown considerable growth over the last decade. Particularly, the COVID-19 pandemic has forced many education systems to switch to e-learning strategies. A major constituent of the online learning experience in both formal and informal learning environments is asynchronous video lessons. These videos have effectively replaced many traditional in-person lectures, and students are now expected to spend increasing amounts of time watching them. While this offers several advantages (e.g., the ability to pause, rewind, and fast forward presentations), there is a need for more empirical studies comparing different video formats augmented with multimodal information. One assumption is that seeing the instructor's face is beneficial to learning; this assumption is so widespread that virtually all online videos include a recording of the instructor's face. However, the advantages of doing so are mixed, with studies finding that it attracts learners' attention at the expense of what is explained (van Wermeskerken et al., 2018) does not significantly improve students' learning (Kizilcec et al., 2014), and increases cognitive load (Kizilcec et al., 2015). There are several reasons why this format is so prevalent. One is technical access (every computer is equipped with a webcam, and there are many softwares that can overlay a person's face onto a screen recording); another one is a penchant to replicate in-person instruction where lecturing plays a predominant role. Because online learning videos are so prevalent, developing better ways of designing them has important practical implications.

In this paper, we explore alternative formats for designing online videos. More specifically, we use theory-driven principles to design videos that can specifically support students' learning. We leverage theories of developmental and social learning, showing that joint visual attention (JVA) plays an important role in learning (Tomasello, 1995). JVA is central to the quality of social interactions, especially when individuals establish a common ground, which has been studied qualitatively (Barron, 2003) and quantitatively (e.g., using eye-tracking; Schneider et al., 2018). The results have inspired various interventions to support collaboration and learning, for example through shared gaze visualizations (SGV). SGVs are either live or prerecorded eye-tracking visualizations that indicate where an individual is looking at. They have been shown to facilitate referencing, increase awareness of others, decrease cognitive load, disambiguate utterances, improve mutual understanding, and reduce verbal effort (for a review, see D'Angelo & Schneider, 2021).

Given this, we tested the benefits of seeing SVGs and the instructor's face on instructional videos. Students (N=52) enrolled in a semester-long course watched weekly videos on using sensor data for education research and filled a weekly quiz to test their learning. Additionally, we collected web-based eye-tracking, emotion, and pose data from participants while they were watching videos. We test the effects of these interventions on learning and investigate whether the interventions impacted joint visual attention, joint emotional responses, and movement synchrony with the instructor. These measures were chosen for theoretical reasons. A first synchrony measure, JVA is an important mechanism for grounding communication (Tomasello, 1995). Thus, we are particularly interested in computing the extent to which students and the instructor align their attention with different interventions using dual eye-tracking data (Schneider, 2020). We believe that this measure would

not only capture whether students are following along but also processing information like the instructor. Second, emotion contagion theories (Parkinson & Simons, 2009) suggest that people tend to mimic each other's affective state in social settings where emotions are shared. In our context, this could signify students are getting engaged when the instructor is showing signs of excitement; or becoming more focused when the instructor is non-verbally communicating that the material is more challenging. In other words, we expect successful learners to share the affective state of the instructor while they are watching the video. Third, body language mimicry has been observed to take place when two individuals are "in tune" (Chartrand & Bargh, 1999) with each other. This could indicate that students are trying to put themselves in the shoes of the instructor by replicating body postures and gestures as a way of processing the material. In summary, each of these measures can mediate learning and contribute to our understanding of how learners comprehend new information from video recordings.

The contributions of this paper are as follows. First, we augment online learning videos with multimodal information (i.e., with the instructor's gaze and/or face) and test the effect of this intervention on learning scores. Second, we explore the usefulness of webcam-based sensing technologies for deriving indicators of students' learning. Most prior work used professional hardware for collecting eye-tracking data; in this project, we investigate whether webcam-based measures can provide relevant metrics for predicting learning. This is an important factor to consider if we are to scale up this kind of approach. Third, we compute measures of joint visual attention, joint emotional response, and joint body movement between students and the instructor and test the mediatory effect of these measures on learning. Finally, we provide some preliminary design principles for augmenting videos with gaze data so that our approach might be replicated in research or practice. We conclude by discussing the implications of our findings and conclude with recommendations for designing online videos.

Research questions

Our main research questions are as follows:

- RQ1: do videos augmented with the instructor's face and / or gaze have a positive effect on students' learning?
- RQ2: do joint visual attention, joint emotional response, or movement synchrony mediate learning?

Methods

Participants

Participants (N=52) were graduate students at a private graduate school of education in the Northeastern region of the U.S. who were admitted to a course on Multimodal Learning Analytics (MMLA). 54% were female and 46% male. Before enrolling in the course, students were asked to complete a brief interest and skills assessment survey. When asked about their level of experience, the following percentage of students indicated having little to no experience using sensor data (81%), data mining techniques (69%), or psychometrics (79%); in contrast, they indicated having some or strong experience (69%) with learning theories such as constructivism or constructionism.

Context

Participants were enrolled in a 13-week course on MMLA in the Fall 2020 semester (for more information on the in-person version of this course, see Schneider, Reilly & Radu, 2020). The course aimed to equip students with knowledge and practice in MMLA methods to collect datasets in various learning environments and analyze them based on theoretical frameworks in the learning sciences.

Due to the pandemic, the course was taught exclusively online. The course structure involved real-time instruction, hands-on projects, and twelve weekly asynchronous videos. For the projects, students collected relevant data via an interactive data collection website with computer vision algorithms, as opposed to physical sensors that were used in previous, in-person offerings of the class (for more information on this tool, see Schneider, Hassan & Sung, 2022). This website allowed students to collect eye-tracking, pose, emotion, physiological and gesture data from videos.

The asynchronous videos taught the following topics: introduction to MMLA, physiological data, eye-tracking parts 1&2, body tracking, supervised machine learning (ML), applied ML, clustering parts 1&2, data preprocessing, experimental and descriptive methods, and Markov Chains. These videos were uploaded to an online learning platform that captured and locally processed participants' webcam feeds as they watched videos to generate data on gaze, emotion (from their facial expressions), and body motion. This data was used both for a mini quantified-self project at the end of the semester and for research purposes. Participants were informed of the purposes of the recordings and had continuous access to their own data on the learning platform. All participants signed a consent form to agree to the data collection process.

Material

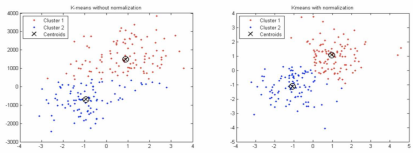
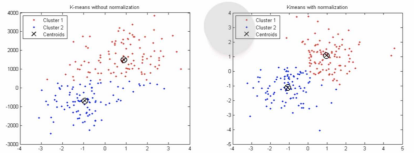
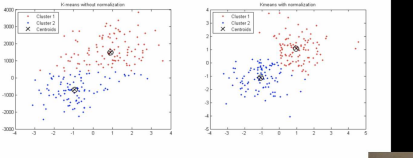
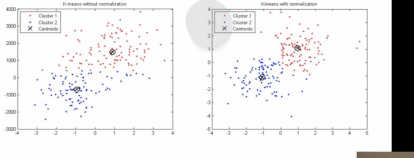
Video lectures were recorded of the instructor teaching MMLA concepts using slide decks. A webcam captured the instructor's face and a Tobii 4C eye-tracker captured his eye movements. Four versions of the videos were created: some of the videos had the instructor's face overlaid on top, and some were augmented with the instructor's gaze (see Table 1).

After each video, participants took a quiz testing their knowledge of the content of the video. There were ten items each week, and included questions about facts, concepts, procedure, and applications in MMLA. For example, in weeks 6-7, where the topic was machine learning, conceptual questions included: "What are potential solutions to a model underfitting the data?"; "What are potential solutions to a model overfitting the data?". Examples of factual questions are: "What of the following algorithms are probabilistic?". Examples of procedural questions are: "What is the first step in the K-means algorithm?". Finally, examples of application questions are: "Which of the following are examples of application of machine learning?" In total, there were 50 factual questions, 35 conceptual questions, 24 procedural questions, and 11 questions about the applications of MMLA.

Design

We used a 2x2 experimental design to test the effect of the instructor's face and gaze on the videos: a quarter of the students saw the raw video; a quarter saw the instructor's face next to the video; a quarter saw the instructor's gaze on the video; and a quarter saw both (Table 1). To test the effect of each intervention, we compare students who saw the videos with and without the instructor's gaze, and students who saw the same videos with and without the instructor's face. Participants were assigned to the same conditions for the entire semester.

Table 1: the 2x2 experimental design used in this study. Each cell contained 13 participants. Each row / column contained 26 participants. The gray circle on the right column indicates the location of the instructor's gaze.

Conditions	No instructor's gaze (N=26)	With instructor's gaze (N=26)
No Instructor's face (N=26)	<p>Data Normalization</p> 	<p>Data Normalization</p> 
With instructor's face (N=26)	<p>Data Normalization</p> 	<p>Data Normalization</p> 

Procedure

Each week, students watched an instructional video on the topics mentioned above. Each video was between 20 and 30 minutes long, and provided factual, procedural, and conceptual information about MMLA as well as about its applications. There was no time limit for watching the video, and students could watch it as many times as they wanted. When they felt ready, students took a 10-question quiz with multiple choice answers. The time limit on the quiz was 15 minutes, to ensure that students used their own notes and understanding of the material to answer the questions. We found that the time limit minimized overreliance on the video. Students had until a specified deadline to complete the quiz, after which the correct answers were released.

Multimodal Measures

We computed several measures from the multimodal data. As a reminder, we captured the instructor's eye gaze (using a Tobii 4C), emotions from facial expressions (using Face-API.js; Mühler, 2021), and pose data (using

PoseNet; Oved, 2018). As students were watching the videos, we collected their gaze (using WebGazer; Papoutsaki et al., 2016), pose (using PoseNet; Oved, 2018), and emotion data (using Face-API.js; Mühler, 2021).

Joint visual attention: for each video frame, we compared the location of the instructor's and students' gaze. We standardized gaze coordinates between 0 and 1 because the video dimensions varied based on students' and the instructor's browser window size and resolution. We then computed the distances between the instructor's coordinates and each students' coordinates and tested different thresholds to determine if they shared the same attentional focus. Because the resolution of webcam-based eye-tracking is less accurate than with dedicated hardware, we considered larger thresholds than traditional eye-tracking studies. While we found similar trends across different threshold values, we considered distances below 0.25 (i.e., a quarter of video size) between two gazes to count as joint visual attention.

Joint emotional response: for each video frame, we looked at the probability distribution of different emotions expressed (i.e., happy, angry, fearful, disgusted, neutral, sad, surprised). We obtained similarity scores by computing the cosine similarity between the emotions expressed by the instructor and each student. Cosine similarity is a common measure of similarity between data expressed as two vectors, measuring the angle between them, and thus representing how similarly two vectors are oriented.

Joint body movement: for each video frame, we looked at the movement generated by the instructor and each student (i.e., displacement of the x,y location of visible upper body joints). We computed the difference in total movement between students and the instructor as a measure of joint body movement, where 0 meant that both moved the same amount, and larger values meant that one person was moving while the other was not.

Data Analysis

In this section, we describe the measures we computed from the data and how we addressed our research questions. For RQ1, we used a between subjects Analysis of Variance (ANOVA) to test the effect of the instructor's gaze and face overlaid on the weekly videos. Learning was measured through students' scores on the quiz questions. For RQ2, we correlated joint visual attention, joint emotional response, and joint body movement with learning scores. Additionally, we built a multiple linear regression model to compare the variance explained by each predictor and tested several mediation models to test whether any of these variables mediated learning based on the participants' experimental condition.

Results

Exploratory Data Analysis

An important step in our data analysis was sanity checking the data collected while students were watching the videos. Since there were important hardware, environmental and individual differences between students, data quality varied widely. For each modality, we computed measures of synchrony with the instructor and plotted the data for the entire semester. We show six exemplar graphs in Figure 1 for the eye-tracking data. Each graph represents a student, and the colors represent the different weeks of the semester. The leftmost two graphs show "healthy" data where there is a sizable dataset for each week. The next two graphs each show 'borderline' data sizes of around five thousand, suggesting that the data may not be comparable to that of other students. The next two graphs show unusable datasets due to missing data. The last rightmost graph shows the number of data points for each student (i.e., each student is represented by one bar). We looked at bar graph to identify "dips" in the data. In the case of the eye-tracking data, we used two thresholds for removing students: when they had less than 5k and 10k data points. We found a stronger correlation with a more conservative threshold (i.e., 10k), which we end up using for our analyses. We used the same approach with all three multimodal datasets.

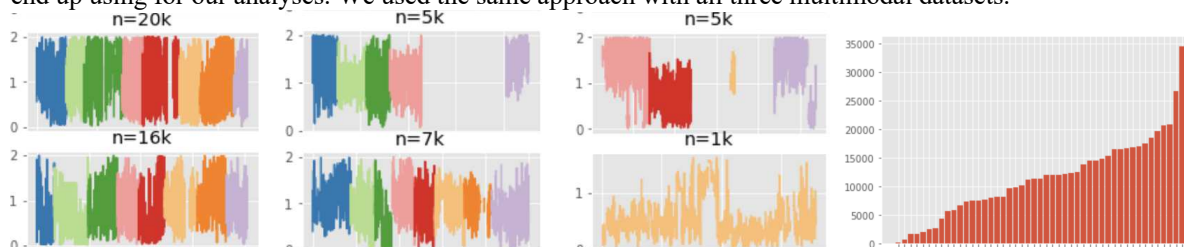


Figure 1: Left side: eye-tracking data for six students (the distance with the instructor's gaze is shown on the y-axis; time is shown on the x-axis; colors represent the different weeks of the semester). Right side: total number of data points collected during the semester (each bar represents a distinct student enrolled in the course).

RQ1: do videos augmented with the instructor's face and / or gaze positively affect students' learning?

To answer the first research question, we aggregated students' quiz scores over the entire semester (see Fig. 2) and tested for significant differences using an ANOVA. We found a significant effect of overlaying the gaze of the instructor on conceptual questions: $F(1,44) = 4.42, p = 0.04$, Cohen's $d = -0.64$ (no-gaze: mean=0.79, SD=0.09; gaze: mean=0.85, SD=0.07). However, we did not find any significant effect of seeing the instructor's face on quiz scores. Additionally, we did not find any significant effect of the instructor's face or gaze on the other learning categories (i.e., facts, procedures, applications of MMLA). Finally, we did not find any significant interaction effect ($F < 1$).

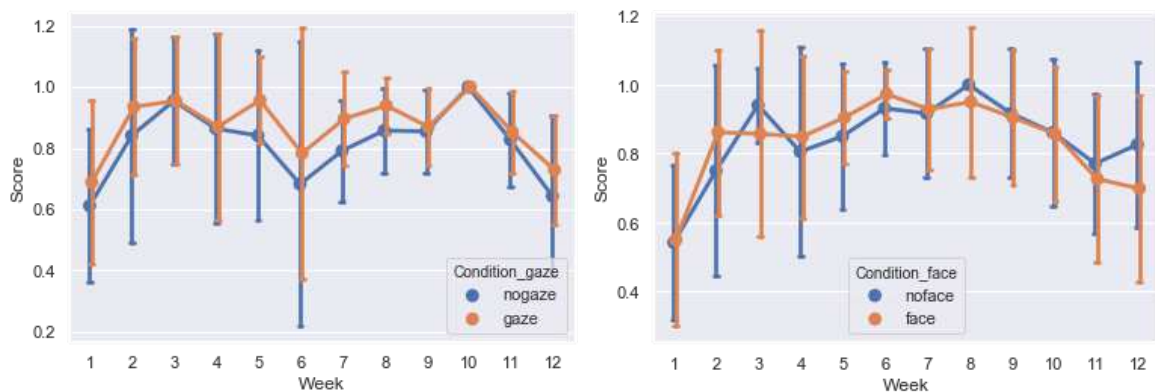


Figure 2: Weekly quiz scores (y axis) of students over the 12 weeks of the semester (x-axis). The left graph shows students who saw the instructor's gaze on the videos compared to a control group. The right graph shows the same information but for the instructor's face. Error bars indicate standard deviations.

RQ2: Do joint visual attention, joint emotional response, or movement synchrony mediate learning?

We conducted three types of analysis to test whether joint visual attention, joint emotional response, or movement synchrony is related to learning outcomes (i.e., quiz scores). First, we correlated these measures with each other and with quiz scores. We then combined them in a multiple regression analysis to understand how much variance they explained. Finally, we tested mediation models to explore if synchrony played a mediatory role between our two video augmentations and learning.

First, we correlated these measures with learning scores. Results are reported in Figure 3. We found that joint visual attention (JVA) was significantly correlated with conceptual scores: $r(24) = 0.530, p = 0.005$, as was joint emotional response (JER). None of the other relationships were significant, which suggests that the three mediators are capturing different constructs and not a general form of engagement with the content (in which case they would significantly correlate with each other).



Figure 3: correlation coefficients between joint visual attention (JVA), joint emotional response (JER), movement synchrony (MOVE) and quiz scores on questions on facts, concepts, procedures, and applications of MMLA. Significant correlations ($p < 0.05$) are highlighted in green.

Second, we built multiple regression models to explore how much variance of learning the synchrony metrics could jointly explain and which ones would stay significant when combined in the same model. We found that the regression model predicted roughly a third of the variance ($R=0.39$, $R^2=0.28$), but only joint visual attention remained significant ($p<0.05$); joint emotional response and movement synchrony became non-significant, which suggest that they do not offer unique explanatory power when combined with JVA.

Third, we tested three mediation models to confirm this hypothesis. The first model (Figure 4) examined the effect of the experimental conditions on conceptual learning, mediated by joint visual attention. The second and third models were similar except that JVA was swapped with joint emotional response and movement synchrony. Because of our small sample size, we used a bootstrapping approach ($n=1000$). Interpretation of the bootstrapped analysis is accomplished by determining whether zero is contained within the 95 percent CIs (thus indicating the lack of significance; Efron & Tibshirani, 1997).

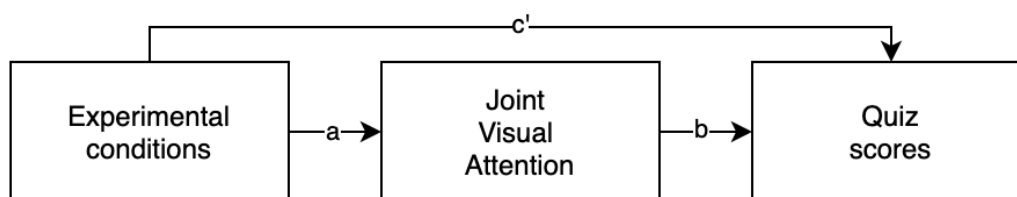


Figure 4: One mediation model tested in this paper (other mediation models replaced joint visual attention with joint emotional response or movement synchrony).

For the first model, the effect of experimental conditions (i.e., the instructor's face or gaze) on joint visual attention was significant ($p<0.01$, CI: [0.0067; 0.037]) as was the effect of JVA on quiz scores ($p<0.01$, CI: [0.21; 1.06]). The total indirect path was significant ($p<0.05$, CI: [0.0024; 0.003]), while the direct path was not ($p=0.85$, CI: [-0.02; 0.023]). For the second model, the effect of joint emotional response had a significant effect on quiz scores ($p<0.05$, CI: [0.05; 0.48]), but there was no other significant effect. For the third model using movement synchrony, we did not find any significant effect.

In summary, the correlation analysis shows that joint visual attention and joint emotional response seem to be associated with higher quiz scores, while movement synchrony does not. It also shows that these measures are not correlated, which suggests that they capture different forms of engagement with the content. Additionally, the multiple regression model and mediation analysis suggest, together, that joint visual attention is the most important mediator for learning. We further discuss these findings below.

Discussion

In this paper, we tested different ways of augmenting learning videos. We compared the effect of seeing the instructor's gaze and face during a semester-long course. Students watched weekly videos and completed a quiz testing their understanding of the material. We found that while adding the face of the instructor to the videos had no effect on students' learning, visualizing eye-tracking data increased conceptual learning (but did not affect other types of learning, such as learning about facts, procedure, or applications). This does not mean that showing the instructor's face is useless; it might influence other aspects of students' learning experience (e.g., their enjoyment, long-term motivation, rapport with instructor, etc.). However, what effect it may have had did not translate into significantly higher quiz scores. The implication of these findings is that it might be more beneficial to add gaze data to learning videos, contrary to common practice of adding faces, if we primarily care about supporting students' conceptual learning.

Additionally, we collected multimodal data from students as they were watching the weekly videos (i.e., gaze, pose and emotion data). We were able to do so because of recent advances in computer-vision based algorithms, such as the ones described by Schneider, Hassan & Sung (2022). We explored synchrony measures between students and the instructor as mediators for learning. We found while none of these measures were correlated with each other, joint visual attention and joint emotional response were significantly and positively correlated with quiz scores. This suggests that students who looked where the instructor was looking at or produced similar emotions (based on their facial expressions) tended to learn more. A multiple regression model showed that only joint visual attention remained significant when all three measures were used to predict learning. Mediation analysis showed that joint visual attention was the only significant mediator between experimental conditions and learning. This suggests that when deciding which multimodal data to collect from students, eye-tracking data (and joint visual attention measures with the instructor) might be the most useful

data source. However, this does not necessarily mean that other synchrony measures should be discarded. It is possible joint emotional response might be useful for capturing rapport with the instructor; movement synchrony might be useful in domains where gestures play an important role in the learning process (e.g., see Son et al., 2018).

In short, the findings of this paper have significant implications for the design of instructional videos. In an age where online learning is becoming ubiquitous, we need empirically based design principles that practitioners can use to support students' learning. Our results suggest that the most widespread format (i.e., where the instructor's face is displayed in a corner of the video) does not affect students' learning; and that there is a need to explore alternative ways of augmenting this kind of media (e.g., by adding eye-tracking data). Additionally, we find that eye-tracking data can help us craft more effective measures of learning by computing indices of joint visual attention with the instructor.

Design considerations

While we found a positive effect of sharing gaze data, we do not believe that it is a silver bullet for making online videos easier to understand. In this section, we describe some (design) considerations for creating effective gaze visualizations. First, there are situations in which they can be useful and others where they likely would not be effective. According to a review on Shared Gaze Visualizations (SGVs; D'Angelo & Schneider, 2021), there is a tension between using gaze to facilitate communication (e.g., by facilitating referencing) and having it distract the viewer (e.g., by displaying fine-grained fixations or saccades). In the design of the videos, the instructor had to consciously minimize scanning behaviors and maximize the use of his gaze as a communication medium. This was not intuitive for the instructor since humans usually do not consciously control their eye movements. While there is time and energy gained by not having to manually annotate every slide with arrows and highlights (which is a long, tedious process), recording eye-tracking data adds cognitive load when recording a lesson. Consequently, working with instructors who know their material well and have been teaching it for several years can facilitate the creation of high-quality learning videos augmented with gaze data.

Limitations

There are several limitations to this study. First, by combining data streams we had to remove participants who did not have enough data. This reduced our sample size and introduced the risk that data may have not been missing at random (e.g., students who were less engaged spent less time on lessons). Second, we acknowledge that webcam-based data collection tools (eye-trackers) are not as accurate as solutions using dedicated hardware. This limits how much we can trust the data, especially across a variety of participants and settings (e.g., race, lighting, hardware). Lastly, related to this point, there are several unknown factors that can compromise the quality of the data. Different students had different hardware, sometimes more affordable laptops that are not suited for processing real-time webcam data. In the future, there is a need to add sanity checks to make sure that the data is reliable (such as the ones described in the "Exploratory Data Analysis" section).

Future Work

In terms of future work, we are planning to replicate these results with another cohort of students. This will increase confidence in the generalizability of our findings. Because recording the videos increases the instructors' cognitive load, we are also interested in developing better tools for recording their gaze (e.g., by changing the type of visualization used, and letting the instructor turn it on and off when necessary). Alternatively, new implications may be found by providing the same functionalities to students, so that they can customize the gaze visualization to their personal preferences (D'Angelo et al., 2019).

Conclusion

In conclusion, we are cautiously optimistic about the potential of SGV for online teaching. While augmenting online videos with eye-tracking data from an instructor is not a silver bullet for making teaching more effective, it can help in some situations (e.g., spatial domains, such as data visualization and data analysis, where there is potential for miscommunication; D'Angelo & Schneider, 2021). Additionally, we found that webcam-based multimodal measures allowed us to design meaningful indicators of learning. We were able to compare different measures of synchrony with the instructor and found that joint visual attention was the most promising indicator of conceptual understanding. These findings contribute to our understanding of the factors that contribute to effective online learning experiences and pave the way for innovative ways of augmenting videos with multimodal information.

References

- Barron, B. (2003). When Smart Groups Fail. *Journal of the Learning Sciences*, 12(3), 307–359.
- Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6), 893–910.
- D’Angelo, S., & Schneider, B. (2021). Shared Gaze Visualizations in Collaborative Interactions: Past, Present and Future. *Interacting with Computers*, 33(2), 115–133.
- D’Angelo, S., Brewer, J., & Gergle, D. (2019). Iris: A tool for designing contextually relevant gaze visualizations. *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 1–5.
- Efron, B., & Tibshirani, R. J. (1997). *An introduction to the bootstrap*. Chapman & Hall.
- Kizilcec, R. F., Bailenson, J. N., & Gomez, C. J. (2015). The Instructor’s Face in Video Instruction: Evidence from Two Large-Scale Field Studies. *Journal of Educational Psychology*, 107(3), 724.
- Kizilcec, R. F., Papadopoulos, K., & Sritanyaratana, L. (2014). Showing face in video instruction: Effects on information retention, visual attention, and affect. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2095–2102. <https://doi.org/10.1145/2556288.2557207>
- Mühler, V. (2021). *Face-api.js* [TypeScript]. <https://github.com/justadudewhohacks/face-api.js> (Original work published 2018)
- Oved, D. (2018, September 27). Real-time Human Pose Estimation in the Browser with TensorFlow.js. *TensorFlow*. <https://medium.com/tensorflow/real-time-human-pose-estimation-in-the-browser-with-tensorflow-js-7dd0bc881cd5>
- Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable Webcam Eye Tracking Using User Interactions. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 3839–3845.
- Parkinson, B., & Simons, G. (2009). Affecting Others: Social Appraisal and Emotion Contagion in Everyday Decision Making. *Personality and Social Psychology Bulletin*, 35(8), 1071–1084.
- Schneider, B. (2020). A Methodology for Capturing Joint Visual Attention Using Mobile Eye-trackers. *Journal of Visualized Experiments*, 155, e60670.
- Schneider, B., Hassan, J., & Sung, G. (2022). Augmenting Social Science Research with Multimodal Data Collection: the EZ-MMLA Toolkit. *Sensors*, 22(2), 568.
- Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2018). Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning*, 13(3), 241–261.
- Schneider, B., Reilly, J., & Radu, I. (2020). Lowering Barriers for Accessing Sensor Data in Education: Lessons Learned from Teaching Multimodal Learning Analytics to Educators. *Journal for STEM Education Research*, 3(1), 91–124.
- Schwartz, D., & Martin, T. (2004). Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. *Cognition and Instruction*, 22(2), 129–184. https://doi.org/10.1207/s1532690xci2202_1
- Son, J. Y., Ramos, P., DeWolf, M., Loftus, W., & Stigler, J. W. (2018). Exploring the practicing-connections hypothesis: Using gesture to support coordination of ideas in understanding a complex statistical concept. *Cognitive Research: Principles and Implications*, 3(1), 1.
- Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Lawrence Erlbaum Associates, Inc.
- van Wermeskerken, M., Ravensbergen, S., & van Gog, T. (2018). Effects of instructor presence in video modeling examples on attention and learning. *Computers in Human Behavior*, 89, 430–438. <https://doi.org/10.1016/j.chb.2017.11.038>