EVA: Exploratory Learning with Virtual Companions Sharing Attention and Context

Scott W. Greenwald, Markus Funk, Luke Loreti, David Mayo and Pattie Maes
MIT Media Lab (75 Amherst St., Cambridge, MA 02139 USA)
{scottgwald, mfunk, lhloreti, dmayo2, pattie}@media.mit.edu

Abstract—Exploratory Learning with Virtual Companions Sharing Attention and Context (EVA) is a concept for mediated teaching and learning that sits at the intersection of exploratory learning, telepresence, and attention awareness. The companion teacher is informed about the attentional state and environment of the learner, and can refer directly to this environment through marking or annotation. To the learner, the companion is virtual—either human or automatic—and, if human, either physically copresent or remote. The content and style of presentation are tailored to the learner’s momentary level of interest or focus, and her attention can be guided to salient environmental elements (e.g., visual) in order to convey desired information. We define a design space for such systems, which applies to learning in Augmented Reality and Virtual Reality, and can be employed as a framework for design and evaluation. We demonstrate this through trials with two proof-of-concept systems, one in AR and one in VR, with a human companion. We conclude that the EVA design space defines a powerful set of systems for learning and finish by presenting guidelines for making such systems maximally effective.

I. INTRODUCTION

In this work we investigate the intersection of exploratory learning, telepresence, and attention awareness. Exploratory learning has been proposed as a model for engaging with learning materials and activities in immersive environments, with the advantages of greater autonomy and personalization, among others [6]. It has been observed that this style of learning is compatible with a variety of learning models. A learner can, for example, fluidly shift from a formal learning activity like watching a lecture, to an informal one like playing with blocks. We propose that sharing a first-person perspective with “companions” who can comment and annotate in real-time may act as a crucial building block for future learning environments.

Face-to-face tutoring is considered a gold standard for learning a wide variety of skills and concepts. Achieving the same level of performance through telepresence would offer the advantages of (1) scalability, greatly expanding who can learn from whom by eliminating the need for physical copresence, (2) enhancing learning in augmented reality, allowing copresent tutoring to occur in places where it wouldn’t be feasible physically, and (3) enhancing learning in virtual reality, where copresence is necessarily mediated. A major factor that makes face-to-face learning so effective is attention awareness—the teacher’s ability to directly observe the student’s momentary focus of attention and level of comprehension, and adjust her presentation accordingly. This includes, for example, reiterating, adding emphasis, or providing more detail. Non-face-to-face forms of communication typically lack some of the non-verbal signals required for this kind of adaptation. As important as attention awareness is the teacher’s ability to refer directly to the shared environment by pointing to or describing objects [14]—a telepresence ability that is also frequently lacking in remote or virtual communication settings. We propose that mediated teaching and learning in immersive environments can go “beyond being there” [11], i.e. match or exceed the effectiveness of face-to-face tutoring for learning, by incorporating in real-time (1) attention awareness and perspective sharing, i.e., capturing a first-person perspective augmented by peripheral gestural and physiological cues, and presenting this to the teacher or peer, and (2) telepresence, giving the teacher the ability to directly make reference to objects in the environment. When combined with exploratory learning, we call this configuration Exploratory Learning with Virtual Companions Sharing Attention and Context (EVA).

At present, a comparatively easy way to provide an effective remote or virtual learning experience is to connect to a learner to a human teacher in real-time. However, the assistive functions of the EVA system may gradually become more autonomous, and ultimately replace the human companion. The term virtual companion describes the teacher role from the perspective of the learner: a provider of timely and relevant information that may or may not be human.

In the present work, we design, implement, and test two minimal proof-of-concept EVA systems. The first is an Augmented Reality (AR) system, which uses a streaming head-mounted video camera to share the learner’s external visual and auditory context. The same video stream conveys rich information about the learner’s momentary focus of attention. The companion interface allows the teacher to place ephemeral markers in the field of view of the learner. In the other, a Virtual Reality (VR) system, a similar visual perspective is shared from learner to teacher, but the environment is virtual. In this case the companion interface allows the teacher to place persistent markers in the environment and draw attention to objects outside the learner’s current field of view. In sections that follow, we first review related work and define the EVA concept in detail. Next, we present a design space for systems that facilitate EVA. Finally, we report on the trials we performed and the results obtained.

II. RELATED WORK

We briefly discuss three areas of related work: learning in AR, collaboration with AR, and learning in VR. In [3], AR was shown to be a viable medium for guiding students
through tasks such as building molecular models – in both co-present and remote teaching scenarios. Numerous works have focused on the use of AR for assistance and collaboration. Work on assistance in AR tends to focus on sophisticated means for object referencing and attention management. Addressing the need to redirect the attention of a user of an AR or VR system to a specific point in a complex scene and/or outside of the current field of view are [2], [3]. Head-mounted display solutions that are similar in spirit have been tried as well [1]. Utilizing active tracking to draw useful indicators onto surfaces with light projections is shown in [9], [12]. The problem of stabilizing such annotations when they are displayed on a tablet or computer screen is addressed for example, by [4] and [7].

Considering next work done in the area of using virtual worlds for learning, [5] notes that virtual worlds are already being used for situated learning, often in the form of online games and simulations with large communities. Currently, these are usually accessed on personal computers and user to user communication is limited to text entry and avatar gestures.

The most similar prior works are [13] and [8]. In [13], the authors investigate the use of AR annotation in collaborative tasks as an augmentation of videoconferencing. They evaluate the system according to the degree of subjective co-presence that is achieved, and conclude that the ability to use a pointer most strongly contributes to a sense of being together. In [8], the authors investigate situated informal learning from a remote expert using mobile phone-based video streaming, still image annotation, and face-to-face communication. Remote attention sharing is shown to be advantageous over physical co-presence in some circumstances, since the latter carries attentional burdens unrelated to the learning content.

III. APPROACH

The basic learning interaction we propose, exploratory learning with virtual companions sharing attention and context, is one where the learner explores an environment in such a way that the focus of his attention is shifted through his own volition, and the teacher or companion provides information or feedback that is relevant to the current focus, sometimes making use of a pointing mechanism to make direct reference to elements of the environment. The companion has access to the external context and attentional state of the learner, and she uses this to judge what input to provide, and at what point in time. Examples of external context information include video-streaming, depth sensing, GPS, and indoor location tracking. Examples of signals of attentional state include eye gaze, EEG, heart rate, electrodental activity, and respiration. We propose the following research question pertaining to such systems that facilitate exploratory learning with a remote companion:

How can giving the remote companion access to the external context and attentional state of the learner improve the learning interaction?

Next, we present two ways of defining and breaking down systems that attempt to answer this question. The first is the 2 × 2 taxonomy of affordances, which provides categories for the “features” systems provide for the learner and companion. Next, we present a number of design dimensions which are choices that can be made in a given case to provide various tradeoffs vis-a-vis the models applied for curriculum, motivation, evaluation, and so on in any given case.

An EVA learning experience involves connecting two or more parties and environments. Each party has some ability to perceive the remote environment(s), and act or express ideas within it. As such, one way of defining or profiling a given system is to enumerate the interface affordances within a 2 × 2 grid of learner and teacher, perception and expression. This taxonomy provides a way to quickly understand what a system does, as a starting point for understanding its design choices. We’ll refer back to it when discussing our proof-of-concept systems.

We now present design dimensions for EVA systems. The first of these is the augmented-virtual spectrum, i.e. the choice between augmented reality, virtual reality, or an in-between mixed reality. Each of these offers different opportunities – augmented reality turns a physical environment into a ready-to-go setting for learning. Virtual reality, on the other hand, allows learners to go places they can’t go physically – such as inside a microscopic cell environment or into outer space. Next is the synchronous-asynchronous spectrum. Systems can be designed to support synchronous or asynchronous communication. Although the EVA concept is centered around real-time feedback, there are cases where it is appropriate (and still “real-time” for feedback to be delivered later) – such as when a teacher comments after observing a performance or demonstration of a performative skill.

Another dimension is the active-passive learning spectrum of content, which refers to the active versus passive roles the companion can take on in determining what is learned or discussed. Consider an application of a video-based mobile communication system where the student walks around a historic city accompanied virtually by the teacher. In the case where the teacher determines the path and the sights to be observed, the interaction has the character of a lesson, where the learner is passive. If, on the other hand, the learner decides where to walk and the teacher volunteers relevant information or answers questions, learner’s role is active.

Next consider the level of personalization that is possible for an EVA system to achieve. When communicating in real-time, the level of personalization may be related to how well the teacher knows the student. If she knows him well, she is likely to be able to adapt presentation very well – knowing how fast to talk, when he is losing focus, and so on. We can think about this in terms of the complexity of the teacher’s mental model of the student. When producing a prerecorded lesson, a teacher necessarily adopts a very generic mental model of the student to whom she is teaching. When tutoring one-on-one the teacher can understand the learning progress of the individual student in a detailed way.

IV. PROOF-OF-CONCEPT SYSTEMS: VR AND AR

We opted to sample the design dimension of augmented versus virtual reality. We built one VR system and one AR
system, both sharing the concept of showing the companion a video stream of the learner’s field of view, and offering the bidirectional expression affordance of placing markers in the shared environment. In addition, both systems allowed the teacher and learner to communicate verbally.

In the VR system, the environment was a photosphere – a continuous panoramic photo surrounding the learner – that could be explored using a VR headset. The teacher could see the entire photosphere as well as what portion the student is currently looking at. He could send markers that were either within the learner’s current field of view or outside of it. In either case we used an attention funnel [2] to guide her gaze to the point in question. Markers initially appeared in a bright color – they were persistent but faded in color intensity after several seconds (in our experiments, even seconds).

In the AR system, the learner was situated in a physical environment to be learned about, and this was captured by a video camera on the front of the head-mounted display device. The markers were not persistent (in experiments, these faded after three seconds), although it would have been possible to use methods such as those demonstrated in [7] to create such markers that stayed stationary with respond to the world.

The learners used a Google Cardboard V2 headset with a Samsung Galaxy Note 5. The teacher used a Microsoft Surface Pro 3 tablet. In the AR trials, the teacher again saw the learner’s field of view exactly as it appeared in the Cardboard headset. In VR trials, the teacher again saw the learner’s field of view, but also had a second window representing their own field of view in the photosphere, which he or she could navigate by altering the tablet’s orientation. These interfaces are shown in Figure 1.

V. EXPLORATORY TRIALS

The first goal of performing trials with the proof-of-concept systems was to evaluate the usefulness of attention and context awareness, i.e. the EVA concept as a whole. The next was to contrast EVA systems along two of our design dimensions – in particular the augmented-virtual spectrum, and the active-passive learning spectrum. Finally we sought to characterize the expression affordance of placing markers in the shared environment, in correlation with variations along the said design dimension. In particular, the two systems allowed us to naturally investigate ephemeral versus persistent markers, and the teacher expression affordance of placing markers either within the current field of view only (AR condition), or also including areas outside the learner’s current field of view (VR only). Our hypotheses were as follows: (1) Attention and context awareness would give learners a sense of copresence and allow companions to quickly respond to the user’s momentary attention, leading to a highly engaging experience. (2) Regarding the augmented-virtual spectrum, the perception of the AR vs. VR systems would be substantially different, given the significant distinction between learning about the real world vs learning about a still VR photo. (3) Regarding the active-passive learning spectrum, learners who were not knowledgeable would prefer passive learning, whereas those who are more knowledgeable prefer the active learning. (4) Regarding expression affordances, persistent markers and marking outside the field of view would be perceived as an advantage for the VR system, and the fact that AR markers do not stay stationary with respect to the environment would be perceived as a disadvantage for that system.

We conducted exploratory trials at an art museum, inviting learner participants (3 male, 4 female, ages 22 to 30), and two expert companion participants (2 female, ages 31 to 34). Learners were guided to particular works of art and directed to converse with the remote expert about each of them. Our expert companions were art historians, and our learners had only a passing familiarity with contemporary art. Each learner participated in a total of four trials – two each with the AR and VR systems (see Table I for details on the conditions). The order of these trials was shuffled for each participant, so that the responses to a particular trial would not be affected by the learning curve of the system. For every participant we designated the first two trials as teacher-driven (passive learning) exercises, where the teacher guided the discussion in a manner similar to a traditional classroom setting, and the second two as inquiry-driven (active learning), where the learner was free to ask questions about whatever piqued his interest.

For quantitative analysis, we considered a repeated measures design with two independent variables: the learning system being used and the required user engagement. As dependent variables, we used the NASA-TLX [10], a measure of subjective cognitive workload, and subjective feedback according to our questionnaire. We counterbalanced the order of the learning systems according to the Balanced Latin Square. For qualitative analysis, at the end of each trial the learner was given a short questionnaire and the TLX survey. Additionally we conducted in-depth interviews with participants subsequent to the trials. At the end of all four trials the learner was interviewed to elicit qualitative feedback about the experience. The teacher was given an extended interview at the end of her entire set of trials.

A. Results

We statistically compared the NASA-TLX score and the results of the questionnaire between the conditions.

Considering the NASA-TLX score between the conditions
Variables Tested

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
<th>Variables Tested</th>
</tr>
</thead>
</table>
| VR Undersea    | 360 image in front of coral reef  | • Target content is 360
• Virtual environment difficult to access physically |
| VR Painting    | 360 image in front of museum painting | • Direct comparison to AR painting |
| AR Painting    | Standing in front of museum painting | • Direct comparison to VR painting |
| AR Sculpture   | Walking around museum sculpture    | • Understanding content requires moving physically |

**TABLE 1**
Trial Conditions for EVA Systems

![Fig. 2. The average NASA-TLX scores according to the conditions used in the study. The error bars depict the standard error.](Image)

The most interest and engagement was perceived by the participants using the AR Sculpture ($M = 6.43$, $SD = 0.73$) and the VR Painting ($M = 6.33$, $SD = 0.47$), followed by the VR Undersea ($M = 6.2$, $SD = 1.17$), and the AR Painting ($M = 5.85$, $SD = 0.99$). Again, a Friedman test did not reveal a significant difference between the conditions ($p > .05$).

**Q4:** The technology platform contributed positively to the quality of the interaction.

Concerning the quality of the interaction, the participants liked the VR Undersea the most ($M = 6.17$, $SD = 0.69$), followed by the VR Painting ($M = 5.33$, $SD = 0.75$), the AR Sculpture ($M = 5.14$, $SD = 1.25$), and the AR Painting ($M = 5.0$, $SD = 1.20$). A Friedman test did not reveal a significant difference between the conditions ($p > .05$).

**Q5:** The technology distracted me from viewing the art piece

Participants found the VR Undersea the least distracting ($M = 2.33$, $SD = 1.37$), followed by the AR Sculpture ($M = 4.14$, $SD = 1.36$) and the VR Painting ($M = 4.17$, $SD = 1.34$), and the AR Painting ($M = 4.43$, $SD = 1.76$). However, a Friedman test did not reveal a significant difference between the conditions ($p > .05$).

**B. Qualitative Results**

Beginning with broad comments on the nature of the experience overall, one participant made a comparison with classroom-based learning:

*That’s my initial gut reaction: as opposed to sitting in a classroom listening to a professor lecture, being able to be in that AR or VR space was more exciting and made me even pay closer attention to details.*

Another participant commented on the sensation of remote presence:

*It definitely felt like there was someone along with me while I was watching. I knew that there wasn’t anyone there, but just the fact that they could see what I was seeing was really helpful, in answering my questions, and also getting the actual tour guide experience, rather than just a voice through a microphone.*

When asked to compare the teacher’s ability to react to their current focus of attention, most participants agreed “I think they were able to react just the same.” One participant commented “It’s nice to be able to say ‘What’s this one?’” One participant offered the caveat that it was difficult for the teacher to distinguish “silence because you’re admiring the piece vs silence because you’re waiting for information.”

These users’ comments are consistent with our hypothesis that the overall EVA concept leads to effective and engaging learning experiences.

Next, contrasting the VR vs AR experiences overall, a striking result was that participants saw very little difference between the two. One participant commented “[AR] felt more authentic because I could walk, move around the space.” However, numerous participants echoed the sentiment “they were about the same,” citing minor technical differences which will be mentioned below. As a result, the vast majority of comments apply to both conditions, referring in general to
the experience of learning from a remote teacher with a shared audiovisual space.

Comparing markers in the VR and AR conditions, we had hypothesized that the persistence of markers in time and space in the VR condition would lead to a significantly different experience. Remarkably, all participants reported not noticing the distinction. After calling attention to the specifics of this distinction during the interview, multiple participants commented that sometimes the markers didn’t disappear or didn’t disappear fast enough. Only one participant commented on the problem of video markers becoming inaccurate during and after movement, although he hadn’t noticed directly that the VR condition did not have this shortcoming. He reported learning to compensate for the problem by waiting until a marker was sent before moving. Asked how fast markers should ideally disappear, one participant stated “Well you really want them to disappear when you don’t care about them anymore,” adding that this would be difficult to infer with the present system apparatus. The fact that participants perceived no difference at all, with a slight preference for having markers disappear faster and more completely, was a surprising result, contrary to our hypothesis.

In addition to the role of marker persistence, participants were asked about the effectiveness and importance of the teacher’s ability to make reference to objects outside their current field of view (system offers support only in VR). As with marker persistence, not a single participant had explicitly noticed this functional distinction between systems. When asked about it, however, every participant reported that the teacher had used this capability, and nearly all found it useful. Pertaining to this, one shortcoming of the system was that it was not easy for the teacher to ascertain in which direction the participant needed to turn to view a given out-of-sight object. Further, even once it was known, it was difficult to communicate to the learner. The two prevailing strategies were to use verbal instructions such as “A little more to the left,” or to use markers in one part of the visual field to indicate a direction. One learner suggested “It would be helpful if she could show me an arrow like in a first-person video game.”

In the EVA section, we drew a distinction between question-answering and lesson-based interactions, and we split our trials half and half between these conditions. The quantitative results showed little difference between these conditions in terms of cognitive load or ease of use, however the qualitative results confirm that this is a meaningful and important distinction. In particular, participants agreed that the proper choice of who should guide the interaction depends both on subject matter and learner preference. Learners emphasized that in environments and contexts they know little about, they prefer to have an expert introduction.

VI. Conclusion

We defined the area of Exploratory Learning with Virtual Companions Sharing Attention and Context, along with a taxonomy for the capabilities such systems may provide, and design dimensions that determine the different ways in which such a system can be useful. We then applied this theoretical framework and explored two proof-of-concept systems through trials in a real-world setting – with art experts teaching learners about artwork at a museum. Even with a very basic signal of learner attention – the head pose in virtual or augmented reality – teachers were able to convey a strong sense of copresence, and accommodate learners’ momentary focus of attention and curiosity. The areas for improvement highlighted by our trials pertain to the teacher’s capabilities: expression (better ways of indicating the relative location of objects) and perception (disambiguating possible reasons for silence), respectively. In order to create a more powerful learning experience, it is the teacher – with the support of the system – who must better understand the state of the learner, and use more powerful tools for expressing contextual (e.g. spatial) information at the right time and in the right way. Given how well a very minimal system performed, enhancing a similar design in either or both of these ways, e.g. through the tracking of attentional state with EEG and/or improved methods for the conveyance of spatial information, appears highly promising for enabling unparalleled learning experiences.

VII. Acknowledgements

This work was supported by gifts from Google and Pearson. We wish to thank Chris Schmandt and Albrecht Schmidt for their valuable feedback.

References