International Society of the Learning Sciences

Traveling Bazaar: Portable Support for Face-to-Face Collaboration

Rosanna Vitiello, Soham Dinesh Tiwari, R. Charles Murray, Carolyn Rosé
rvitiell@andrew.cmu.edu, sohamdit@andrew.cmu.edu, rcmurray@andrew.cmu.edu, cprose@cs.cmu.edu
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract: For nearly two decades, conversational agents have been used to structure group interactions in online chat-based environments. More recently, this form of dynamic support for collaborative learning has been extended to physical spaces using a combination of multimodal sensing technologies and instrumentation installed within a physical space. This demo extends the reach of dynamic support for collaboration still further through an application of what has recently been termed on-device machine learning, which enables a portable form of multimodal detection to trigger real-time responses.

Vision and Learning Application

As the theoretical foundation for collaboration support has advanced for nearly two decades (Fischer et al., 2007), so has work towards enabling this support to respond dynamically to real time cues from ongoing collaboration. Dynamic collaboration support requires sensing technologies to monitor collaboration within an environment in real time (Adamson et al., 2014; Deiglmayr & Spada, 2010; Rummel et al., 2008). Sensors may capture eye motion, physical motion, facial expressions, touch, speech, or video (Cukurova et al., 2020). These sensors communicate detection of key events within a collaboration to a server that collects these signals and maintains an up-to-date representation of the collaboration such that key events and states can be detected, which then triggers a just-in-time intervention to be introduced into the collaboration environment. The earliest forms of dynamic collaboration support operated within a fully online chat space. In a chat space, sensing technologies can be realized through text classification or other text processing technologies. The innovative technology presented in this demo is a portable form of dynamic collaboration support. This demo represents a landmark in enabling efficient and portable real-time detection of and response to events within a video stream as a proof-of-concept, which paves the way towards expansion to multiple modalities in future work, building on the proposed plug-and-play architecture.

Beyond the technical challenge of efficient real time processing of space intensive data streams like video, support for collaboration requires an architecture that is designed to enable accumulation of meaningful signals from collaboration over time, to keep track of the collaboration state as it changes, and to integrate those two things so that interventions over time will be properly situated within the collaboration. For that purpose, we build on the foundation of the Bazaar architecture (Adamson et al., 2014), which has provided dynamic support for collaborative learning in classroom studies at the middle school, high school, and college level as well as in Massive Open Online Courses (MOOCs) (Rosé & Ferschke, 2016). Initially, Bazaar-supported collaborations were housed specifically in online chat spaces. The restriction of housing CSCL activities in fully online chat spaces allows less expressivity between students than is found in the now ubiquitous online video conferencing environments, like Zoom. Instrumented rooms enable face-to-face collaboration, which might be more desirable in the workplace or in classrooms where the learning activities involve physical work, such as science labs, physical education, sports, or making (Wang et al., 2020). There are, however, scalability issues with instrumented rooms. Furthermore, when students are working in a team over an extended time, many impromptu collaborative work sessions occur when students gather informally in hallway chats or over meals. A portable setup, such as presented in this demo, would make it possible to monitor and intervene more effectively in impromptu collaboration settings. We thus refer to our demo infrastructure as Traveling Bazaar.

Demo Scenario and Overview

In the demo scenario, students work together to configure a robot arm to move blocks from one side of a table and arrange them on another side of the table. Students must collaboratively construct a plan for the robot to enact, and then program it using the robot's simple instruction language. The dynamic collaborative support agent acts as a group discussion facilitator. It gives instructions, asks for feedback, identifies which student has raised their hand to speak, and then calls on them. It uses Accountable Talk Facilitation to draw out reasoning from students and direct them to interact with one another. For example, "Which block will be easiest to program the robot to grab first? Who would like to make a proposal?" Student A raises their hand. "Student A, what is your proposal?" Student A responds. "Would anyone like to challenge that proposal?"

Technical Description

The key challenge to overcome towards enabling portable multimodal learning analytics involving space-intensive data, like video, is that it is extremely slow and potentially expensive to port wholesale in its raw form over available wifi in many locations. The key is therefore to be able to collect and compress the space-intensive data into a compact and informative representation that can be communicated and then used very efficiently by a collaboration server running in a different location. In our demo, this in-situ monitoring of face-to-face collaboration is realized through a RealSense D435i depth camera that directly feeds a video stream to an on-device machine learning model housed on a Jetson Nano. In future work, other peripherals can be connected to the Jetson Nano so that other modalities can be included, as they have been in stationary multimodal collaboration support infrastructures such as the Smart Office Space, which makes use of input from a microphone, a keyboard, and a Kinect video and motion-sensing device (Wang et al., 2020).

In our Traveling Bazaar infrastructure, the information from the on-device model running on the Jetson Nano is then sent to a server, which communicates to a virtual human agent on a web browser that can be run on a mobile phone or laptop. The resource is designed as a transportable tool to administer and monitor collaboration between students. Altogether, the Jetson Nano, camera, and web-browser device are designed to weigh only a few pounds and to be stored easily in a handbag. We envision that students can carry this tool with them and easily arrange the agent and camera to help monitor and support their collaboration on the go. Figure 1 visualizes this set-up while Figure 2 details the architecture of the system.

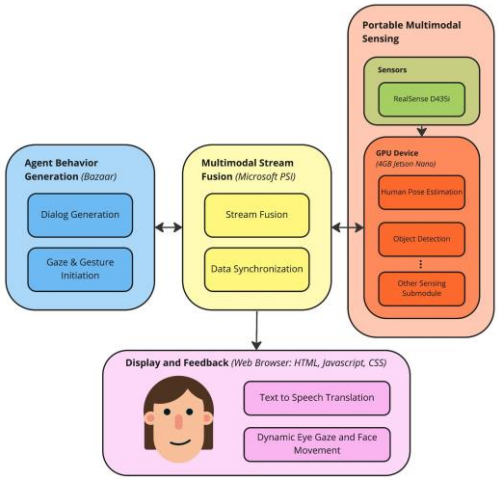
The key technological advancement of this work is the portable multimodal sensing module enabled by the Jetson Nano. By sending sensor streams directly to a machine learning model on-device, the infrastructure avoids the expense of sending large video files across wifi to process on a server. Instead, the video stream is processed locally on-device and compressed into a compact message to send to the server, providing faster throughput for real-time monitoring and better timing of dynamic support. Moreover, the portability of the Jetson Nano allows this system to be mobile, less constrained, and more widely accessible than prior work that provided dynamic agent support in an instrumented room (Wang et al., 2020).

Figure 1

Traveling Bazaar configuration on a desk

Figure 2

Traveling Bazaar architecture and information flow



Information Flow

The flow of data through the Traveling Bazaar system is displayed in Figure 2. Information flow begins first with color and depth frame data captured by camera sensors connected to the Jetson Nano device. This captured data is processed directly on-device: the frames are translated into pose estimation, object detection, and person location information via optimized computer vision models. The computer vision information is compressed into a compact message and sent to the Platform for Situated Intelligence (PSI) (Bohus et al., 2017). This multimodal sensing information, such as raised-hand events or other collaboration events, including pose and location coordinates, may then be sent either directly to the in-browser virtual agent or passed to Bazaar for further event processing. With this sensing information, Bazaar can actively listen to student activity to decide the best course of action for the agent given the current situation. Bazaar passes instructions for the agent, such as verbal text and

facial gestures, back to PSI, which then forwards the instructions to the in-browser virtual agent. The agent translates verbal text to audible speech and displays the non-verbal actions in a browser on a mobile phone or laptop.

Portable multimodal data sensing

In this module, video data is collected directly and locally processed on-device via a 4GB Jetson Nano. Running computer vision machine learning models on the portable GPU device allows for faster processing by avoiding sending large video files to a GPU on a server. Moreover, the portability of the sensors and Jetson Nano GPU enables the system to be mobile. To calculate key body points and objects, we use an optimized ResNet (He et al., 2016) pose estimation model and object detection model provided by NVIDIA's Jetson Inference library. For location verification, we use the Intel RealSense depth-sensing camera to estimate a 3D coordinate position for every detected person. These models process video frames at a rate of 7-10 frames per second on-device, which is sufficient for real-time monitoring and dynamic support. Additionally, this module may also be used standalone for post-hoc multimodal learning analysis on previously collected video of collaborative interactions.

Multimodal stream fusion

To coordinate data from multiple streams and modules, we use the PSI open-source framework (Bohus et al., 2017), commonly used to alleviate engineering problems faced when developing multimodal applications with complex multimodal sensor streams and component AI technologies. PSI associates timestamps with every received event and data stream. This eases coordination for streams and events that run in parallel, such as analyses or recognizers running separately on audio or visual data. Consequently, PSI can synchronize and order events from all data streams so that Bazaar can make informed and time-sensitive decisions for appropriate actions by the virtual agent. In addition to its convenient stream coordination capabilities, PSI logs all collected data in chronological order, which may be used for playback as well as statistical and offline machine learning analysis.

Agent behavior generation

In order to offer dialogue-based support, we use the Bazaar toolkit (Adamson et al., 2014), which has been popularly used to provide dynamic support to students in online collaborative discussions. Bazaar conversational agents use an approach called Academically Productive Talk (APT) that encourages students to elaborate on their reasoning and extend on the deliberation of their peers in group discussion to better enable the creation of group knowledge. The Bazaar toolkit's API provides flexibility to author any learning activities across a variety of topics as desired. Due to its versatility, Bazaar has been used to conduct a multitude of studies in both text-based and multimodal environments. In this context, we use Bazaar in a multimodal setting with adjustable and mobile surroundings.

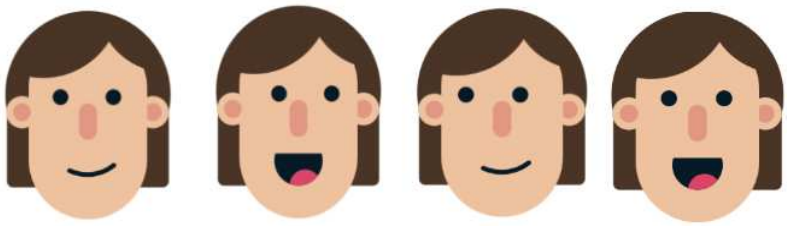
Using event and sensing information from PSI, Bazaar monitors student activity and decides the most appropriate action for the agent given the stage of the collaboration. For example, when Bazaar asks for students to raise their hand to answer a question, Bazaar probes PSI for a raise-hand event and the student's location so that it can instruct the agent to look at the student and call on them to offer their thoughts. As such, Bazaar's responses can be tailored to any context given that sufficient sensing information is provided.

Front end display, feedback, and communication to users

To enable communication of a virtual agent to users, the front end displays non-verbal actions and communicates via verbal text-to-speech in-browser on a mobile phone or computer. The front end consists of two main components: the virtual agent displayed in-browser and its communication with PSI via a Python web server intermediary. We designed the virtual agent SVG using Figma and enabled animation using vanilla client-side JavaScript (JS) in HTML, which adds support for dynamic eye and mouth movement during speech (see Figure 3). The virtual agent also provides text-to-speech support using inbuilt browser capabilities. The browser agent has been tested on Chrome Windows 10 and Safari IOS and is designed to handle a variety of popular configurations on Android and Mac OS.

To establish communication between PSI and the virtual agent, we use an intermediary Python web server to relay information between the front end and PSI. The Python web server is necessary because client-side JS requires WebSockets for communication, which PSI does not currently support. PSI forwards instructions from Bazaar to direct the virtual agent's eye gaze and mouth movement and its speech during group collaboration. For example, PSI sends instructions to direct the agent to look right or left in order to make eye contact while conversing with users.

Figure 3
Example animation configurations of virtual agent



Current Directions

This proof-of-concept demo for portable real-time detection of and response to events within a video stream enables future work to provide support for additional modalities and sensing capabilities. The Traveling Bazaar architecture allows both flexibility in conversational agent authorship through Bazaar and the ability to create additional support for more sensors and machine learning sensing models on the Jetson Nano. Using this plug-and-play system, current work on Traveling Bazaar aims to improve dynamic collaboration support in several ways. To enhance the portable multimodal sensing module, we aim to add additional audio sensor and speech recognizer support and to provide further sensing models on the Jetson Nano in order to better understand student activity in collaboration. Furthermore, in order to build a more robust agent that can operate with looser constraints, we hope to provide a calibration step that maps the working environment so that students can place the tool anywhere in the room. By refining work on this demo, Traveling Bazaar contributes a promising avenue for accessible and portable dynamic support for face-to-face collaboration.

References

Adamson, D., Dyke, G., Jang, H., & Rosé, C. P. (2014). Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of Artificial Intelligence in Education*, 24(1), 92-124.

Bohus, D., Andrist, S., & Jalobeanu, M. (2017). Rapid development of multimodal interactive systems: A demonstration of platform for situated intelligence. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 493–494.

Cukurova, M., Giannakos, M., & Martinez-Maldonado, R. (2020). The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology*, *51*(5), 1441-1449.

Deiglmayr, A., & Spada, H. (2010). Developing adaptive collaboration support: The example of an effective training for collaborative inferences. *Educational Psychology Review*, *22*, 103-113.

Fischer, F., Kollar, I., Mandl, H., & Haake, J. M. (Eds.). (2007). *Scripting computer-supported collaborative learning: Cognitive, computational and educational perspectives* (Vol. 6). Springer Science & Business Media.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Rosé, C. P., & Ferschke, O. (2016). Technology support for discussion based learning: From computer supported collaborative learning to the future of massive open online courses. *International Journal of Artificial Intelligence in Education*, *26*, 660-678.

Rummel, N., Weinberger, A., Wecker, C., Fischer, F., Meier, A., Voyiatzaki, E., Kahrimanis, G., Spada, H., Avouris, N., Walker, E. and Koedinger, K., 2008, June). New challenges in CSCL: towards adaptive script support. In *Proceedings of the 8th international conference on International conference for the learning sciences-Volume 3* (pp. 338-345).

Wang, Y., Murray, R. C., Bao, H., & Rosé, C. (2020, July). Agent-based dynamic collaboration support in a smart office space. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 257-260).

Acknowledgments

This research was funded in part by NSF grants DUE 2100401 and DRL 1949110.