Analyzing Learning with Speech Analytics and Computer Vision Methods: Technologies, Principles, and Ethics

Elizabeth Dyer, Middle Tennessee State University, edyer@mtsu.edu
Cynthia D'Angelo, University of Illinois, Urbana Champaign, cdangelo@illinois.edu
Nigel Bosch, University of Illinois, Urbana Champaign, pnb@illinois.edu
Stina Krist, University of Illinois, Urbana Champaign, ckrist@illinois.edu
Joshua Rosenberg, University of Tennessee, Knoxville, jmrosenberg@utk.edu

Theme, goals, and expected outcomes or contributions

We propose a half-day workshop about video and audio data collection methods that allow researchers to effectively use emerging analytical methods that leverage speech analytics and computer vision techniques, in combination with human-focused analysis (e.g. qualitative analysis). The main goals for participants attending the workshop are to:

1. Become familiar with innovative computational methods (e.g., computer vision and speech analytics) that can be used directly with audio and video data, and consider how computational methods can be used with human-focused analysis to develop new theory in the learning sciences.
2. Understand which features of audio and video data have a large influence on whether computational methods can be applied successfully.
3. Develop principles and strategies for collecting audio and video data in learning environments that increases the successful application of computational methods, including equipment positioning, recording formats and codecs, and equipment features or specifications.
4. Consider ethical implications of using innovative computational methods, both in terms of ethics of conducting research with these methods and potential uses of these methods for education practice and policy.
5. Contribute to a collective methodological research agenda and goals for future development of existing computer vision and speech analytics methods for learning sciences research.

The primary outcome of the workshop is for participants to be able to make informed decisions about collecting audio and video data of learning that will make it possible to use computational methods in analysis. The session will also produce a methodological research agenda for improving the computational methods in their applications to research questions and data used in the learning sciences.

Relevance to the field and conference theme

The learning sciences has a long history of using video and audio to examine processes of human interaction that unfold over time (Goldman, Zahn, & Derry, 2014). Video enables researchers to capture longitudinal data on processes that unfold over time and at multiple time scales, leading to analyses that consider connections between micro- and macro-level phenomena. Video and audio data, in conjunction with multi-modal records of activity, continue to be a central data source in learning sciences, particularly for examining social processes of learning and development.

As capturing and storing video data becomes increasingly accessible and cost-effective, video is emerging as a dominant source of "big data" in the social sciences. Large video databases allow researchers to see social phenomena first hand and provide both breadth in timespan (footage that spans weeks or months of activity) and detail (a rich moment-to-moment interactional and spatial record; Goldman et al., 2014).

Despite the promise and opportunity to use video data in new ways, analytic methods and tools for video have lagged behind innovations in data collection methods, data analytics, and visualization. Video research often relies on processes developed for text-based data (e.g., creating and analyzing transcripts), essentially hiding the temporal and visuospatial dimensions of the data. These traditional analysis methods limit the ability to apply humans' sophisticated visual processing, such as tracking movement over time or seeing relationships in spatially-aligned data points.

Recent advances in computational methods (e.g., computer vision, speech analytics) provide exciting new opportunities to improve the analysis of learning in video and audio data, particularly in large datasets. Some examples of these methods include automated detection of body positioning, emotion, gaze, collaboration, tone, speakers, and prosody. However, because these methods rely on computational power, which differs from human interpretive power, they have different requirements of the data quality and quantity. In the case of speech

analytics, computers are able to do many things, but if the wrong type of audio data is collected (e.g., using a lapel mic to try to capture whole class audio), the computational methods are limited in how well they can interpret the data (Richey, D'Angelo, Alozie, Bratt, & Shriberg, 2016). Additionally, speech analytics methods benefit from high resolution audio that may be almost indistinguishable for a human. With computer vision methods, it can be difficult for computers to identify a person over time if they leave the frame at some point in the video (Wu et al., 2019). Each of these examples represents concerns relating to the quality of the audio and video data that are unique to their use with computational methods. As a result, there are new considerations and principles for collecting video and audio data that can be successfully used with new computational methods.

We, along with other scholars, argue that computational methods are most powerful when integrated with human-conducted analysis and decision-making (Baker, 2016; Berland, Baker, & Blikstein, 2014; Nelson, 2017). These arguments come from a concern over losing the richness and complexity inherent in learning for the sake of convenience and scalability. Additionally, they recognize that humans and computers often have different analytical strengths. For example, Baker (2016) argues for relying on the computational system simply for reporting relevant, low-inference information and patterns, which humans can use for higher-inference analysis to guide future action. We believe that these computational methods provide an opportunity for greater methodological interdisciplinary when they are used in a methodological framework that combines computer- and human-focused analysis, such as computational grounded theory (Nelson, 2017).

Theoretical background

Learning environments are complex social systems in which learning—shifts in knowledge, its collective use, and the related patterns of interaction that demonstrate knowledge development in use—is an emergent outcome. Developing theory about learning requires understanding how interactive (i.e., social and spatial) aspects of classrooms are integral parts of student learning. For example, aspects such as the nature of collaboration, use of gesture and embodiment, the nuances of discursive tone and prosody, and student positional identities are important for understanding learning (Esmonde, 2009; Roth, 2001). This work has demonstrated the need for research methodologies to capture and represent the complexity and nuance in social and spatial aspects of learning. As such, researchers have consistently argued that video and audio data are especially well-suited to capture the visuospatial and acoustic features of interactive processes.

Current research methodologies require Herculean efforts to conduct analyses that simultaneously attend to complexity *and* nuance at a large scale, especially with video and audio data. There are strong qualitative traditions that actively attend to—and even prioritize—visuospatial and/or acoustic features (e.g., Jordan & Henderson, 1995), but these methods are incredibly arduous and time-consuming, making it all-but-impossible to carry out more than a few rich case studies. For example, qualitative studies that look across multiple contexts (e.g., comparing across 100 classrooms) and long time scales (e.g., tracking changes across multiple school years) are incredibly rare. In practice, video data are often reduced to text: transcripts of words spoken, which sometimes include meta-discursive markers or descriptions of gesture. This is a problem, as text is a poor representational form for capturing and communicating visual, spatial, and acoustic dynamics. However, the small repertoire of alternative representational practices for analysis reflected in the literature (e.g., multimodal transcription; Bezemer & Mavers, 2011) are incredibly time-consuming. These challenges to analyzing visuospatial and acoustic aspects of video are partly due to human limitations: people cannot simultaneously attend to all the multimodal dimensions of video and audio data systematically or recognize patterns in these dimensions, even with small data corpuses or a focused microanalysis. As a consequence of these challenges, we need new methodologies for analyzing the social and visuospatial dimensions of learning in video and audio data, especially with the potential to do so at scale.

Computational methods have shown promise for modeling and investigating complex phenomena with large corpora of data, including educational phenomena (Berland et al., 2014). For example, analytic techniques such as vector-space models, topic models, and deep learning/neural networks have all been applied meaningfully to educational research. Importantly, advances in applying these models to educational data sources show their potential for increasing coding efficiency (e.g., Liu et al., 2016), making analysis of large datasets more feasible; and they can be used to detect change over time (e.g., Sherin, 2013), making longitudinal analyses more feasible.

Recent advances in computer vision, coupled with existing speech analytics methods, make it feasible to identify theoretically and practically important features from video in ways that preserve the complexity and nuance that draws educational researchers to audiovisual data—particularly with respect to visuospatial and acoustic features of learning. As an example, computer vision techniques have advanced to the extent that it is possible to use 2D cameras to identify body positioning for multiple people in real time (*OpenPose*; Cao et al., 2017). OpenPose estimates the position of up to 135 key skeletal points (e.g., location of each ankle, finger, top of head, etc.) for individuals in still images and videos. It is robust to partial occlusion, which is key for