Multimodal Data Capture Evaluation Guide

Capturing multimodal data that can effectively be used for learning analytics research can be challenging. In this guide, we provide some suggestions around the types and quality of data needed to conduct meaningful analyses. The suggestions for the modalities and data quality are based on prior work. While we have made every effort to include a rich set of modalities here, authors are encouraged to utilize additional modalities as they see fit and consider sharing their findings with the larger community.

This document also serves to provide guidelines on how grand challenge submissions will be evaluated. In particular, submissions will be rated along several possible optimization dimensions. In describing a given system, authors should clearly identify the dimensions that they were intended to optimize. A description of each dimension is featured below.

Cost
The total cost (both in economic and human resources) is an important variable that determines the feasibility of replicating and applying the capture design. Authors should provide a rough estimate of the total costs of procuring, installing and operating the proposed solution.

Scalability
Learning environments vary widely in terms of size, from small working groups involving just two students to large auditoria with hundreds of them, authors should report on how feasible is for their design to adapt to larger or smaller settings.

Flexibility
Learning environments also vary widely in terms of disposition and arrangement, from traditional lecture settings with the instructor on front and sitting students to more active modes with workgroups of students sitting around tables and a group of instructors roaming among them. The authors should report how feasible it is for their design to adapt to different classroom arrangements.

Intrusiveness
Ideally, multimodal recording solutions for the classroom should be effectively transparent for both the instructor and the students. In reality, different solutions create some level of intrusion (cameras pointing at participants, using special devices to conduct activities, wearing sensor or markers, etc.). As the natural behaviour in the classroom could be altered by the recording setup, the authors should provide a small reflection on the intrusiveness of their designs.
Data Quality

Multimodal Fusion
Across all of the data there is an expectation that each data stream can be resolved to a specific individual (or, in some cases, a pair or triad of students), and to a properly synchronized point in time. Several audio/video, and software based approaches can be used for doing data synchronization, but this information should be clearly reflected in the submission text, as synchronization dramatically impacts data quality.

Skeletal Tracking (and Motion)
Skeletal tracking captured using the Kinect sensor should be frontal. For alternative motion data capture tools (e.g., 2D/3D computer vision algorithms, motion capture with markers) see the appropriate guidelines on collecting high quality data as described in their documentation.

Audio
Prior work in speech and audio process has established a minimum frequency of 8Hz for speech recognition, and higher frequencies, between 12 Hz and 24 Hz, for conducting prosodic and spectral analysis. Additionally, in order to identify various learning relevant constructs (e.g. collaboration quality, voice quality and individual development) the audio captured should have the capability of being resolved to an individual speaker.

Video
One reason for capturing video data is to do facial expression analysis and head pose analysis. Facial expression analysis using computer-automated techniques typically requires a minimum ear to ear width of 64 pixels. A minimum size of 128 x 96 pixels is recommended, as some facial analysis software requires larger images in order to increase accuracy of detection. Additionally, a frontal view is required for being able to conduct facial expression and head pose analysis. However, the data capture environment is free to determine the appropriate frequency of frontal face capture, provided this is a modality of interest.