**Data Management Plan**

**1. Expected data and format**

The data expected from the proposed project are as follows:

1. **Raw image and video data** will be recorded and stored in a current standard formats (e.g. JPG for images and MPEG for videos).

2. **Annotations** on image and video data will be stored in the project MySQL database.

3. **Software programs** that analyze images and video and implement the social network estimation functions will be written in standard programming languages (e.g. C/C++, Matlab, and Python).

4. **Student demographic, registration, and performance data** (e.g., grades) will be recorded in the project MySQL database, and transformed into other appropriate formats such as XML or CSV for accessibility to other software.

5. **Specifications of procedures to be followed and numerical results of performance for benchmark evaluations** will be presented in PDF, XML, and CSV data formats.

6. **Conference papers, journal articles, theses, and reports** written by the students and faculty participating in the project will be in standard PDF format.

7. **Presentations and posters** created by the students and faculty participating in the project will be in standard PDF format.

8. **Presentation slides, Instructional tutorials and reports, and problem sets** associated to curricular materials for computer vision education will be in PDF format.

**2. Data collection**

For the project of student social network estimation from visual observations, we will use data from an audio-visual recording system that has already been built to enable large-scale digitization of student interactions in a Harvard College lecture hall. The system records audio and video from approximately one hundred students during each lecture.

We achieve a frame-rate of 14fps from each of these six cameras, and the combined data-rate is approximately 8.7 MB/s producing about 51 GB of video data from a ninety-minute lecture. The audio sub-system is an innovative design consisting 48 omnidirectional boundary microphones mounted inconspicuously among the seats, along with three 16-channel digitization boxes mounted below the floor of the lecture hall. Each of the 48 audio signals is sampled at 48 kHz with 16-bit quantization, producing about 26 GB of audio data from a ninety-minute lecture, though in this proposed research audio recording will not be employed.

Student demographic, registration, and performance data is also acquired from course registration database hosted by Harvard University, surveys administered prior to the video collection period, students’ per-class registrations about their seating in the classroom through a web-based and mobile-device-based response application, as well as their completed assignments, attendance, and exams agglomerated through the teaching personnel.

Previous data collection and storage with this system has been carefully coordinated with, and approved by, the Institutional Review Board at Harvard University.

In addition to these datasets, smaller video datasets of social interactions between informed actors will be collected. These will be annotated with textual information about interactions, and encoded in standard video formats (e.g., AVI, MOV).

**3. Data storage, archiving, and availability**

All of the data is recorded by a secure server with 48TB of direct attached storage in a redundant RAID configuration, and this is network-attached to an additional server for additional computation. The servers and the data are securely managed by the Faculty of Arts and Sciences computing group at Harvard University, and they are only accessible to a handful of authorized users (members of the project team). Raw data stored on the server will be subject to strict identity credential verification (e.g., secure VPN), and digitized research data will be housed in password-protected files.

Sharing of data with other researchers will happen in accordance with both the NSF Data-Sharing Guidelines and human-subject research restrictions. If and when privacy-preserving technology allows, privacy-preserving editions of the classroom data will be made publicly available. The smaller staged datasets will be made available for direct download through the PI’s websites, along with clear documentation for use.

Interim and final versions of all source main software programs, including source code, will be archived permanently and made freely available from the PIs websites.

Knowledge from this research will be disseminated through journal publications and conference presentations and in workshops. Preprint versions of all manuscripts will be posted on the PIs’ personal websites, as will all theses and reports related to this project. Final versions of published papers will be archived by the publisher, such as IEEE. Digital materials for presentation and posters will be on the PIs personal websites.

Instructional documents, and problem sets, and software programs related to computer vision education will be publicly shared via our custom project website.