Volume 5 | Issue 1 Article 22

2016

How to IMPRESS: Coordinating a Large Video Data Set for a Collaborative Project

Noah-Kee Marks
University of Pittsburgh, noah.marks1@gmail.com

Rita Dawod
DePaul University, ritdaw3@gmail.com

Acknowledgements
This research was made possible by the IMPRESS team at Rochester Institute of Technology. This research was also partially funded by the Research Experience for Undergraduates (REU) at Kansas State University and the Undergraduate Summer Research Program (USRP) at DePaul University.

How to IMPRESS: Coordinating a Large Video Data Set for a Collaborative Project

Noah-Kee Marks*
University of Pittsburgh

Rita Dawod*
Department of Physics

Mary Bridget Kustusch, PhD; Faculty Advisor
Department of Physics

ABSTRACT Though many different research methods involve mass quantities of video/audio data, there is little discussion of best practices for organization, especially when the research is collaborative. The guidelines we provide here were created while conducting educational research on the IMPRESS project, an integrated metacognitive program for at risk STEM majors at Rochester Institute of Technology (RIT). Our hope is to encourage large-scale, collaborative research of qualitative video data by using our project as an example and providing enough information for readers to make a judgment on the efficacy of this process for their own projects.

INTRODUCTION

The utility and diversity of video data has been well documented (Engle, Conant, and Greeno). Many different methodologies can be applied to the data, and various types of claims can be sought from it. The current state of technology allows researchers to collect and archive hours upon hours of video files, and this practice is becoming increasingly common. We aim to show an effective method for organizing large quantities of qualitative video data. The guidelines we are providing were created while working on the IMPRESS (Integrating Metacognitive Practices and Research to Ensure Student Success) research project ("Welcome to IMPRESS," 2015). A description of this specific example of a collaborative video research project provides the information necessary for readers to make a judgment on the efficacy of this process for their own projects. Instead of creating a template of the type of project our system applies to, we will discuss details of IMPRESS that speak to the structure of its data. Highlighting details that the research team actually worked on allows us to better describe

*njm66@pitt.edu, ritdaw3@gmail.com
Research Completed in Summer 2015

our guidelines in a manner that other researchers will find fruitful.

IMPRESS is a project at the Rochester Institute of Technology (RIT) that aims to improve the persistence in STEM majors among First Generation and Deaf/Hard of Hearing. RIT enrolls a high number of each group of students compared to other post-secondary four-year institutions. These two groups, on average, are less likely to graduate than the general student population (Franklin). One of the measurable goals of IMPRESS is to eliminate this gap in student success. An activity that attempts to improve retention is a two-week summer bridge program in August. Roughly 20 incoming freshman students work on open-ended experiments in small groups, have quiet reflection time, and participate in whole group discussions. The specific aims of the bridge program are to improve students' metacognitive skills and to develop a community of learners; both of which have been shown to improve persistence in STEM majors (Mytkowicz, Gross, Steinberg).

Data Quantity and Organization

A video data set is often obtained from natural settings (*i.e*. not staged), and thus there are ample opportunities for it to be acquired. This, accompanied with effective collection and storage processes, makes it very easy to accrue hundreds of gigabytes and thousands of video files. Additionally, data collection and storage tools are relatively cheap; it is only necessary to have researchers, cameras, and hard drives.

The IMPRESS project was not specifically designed for research; however, the video data collected was purposed for research. Almost all of the data was video data. The bridge program had nine days of data collection with about five hours of video every day. We used four cameras to simultaneously record from different vantage points.

Because of program specific circumstances, like activities outside of the lab the students were usually in, we did not always get five hours each day, nor did we always use all four of the cameras. Nonetheless, we accumulated about 150 hours of total video. The cameras broke up videos into five-minute segments, giving us about 1800 video files. Each segment took up about 290 megabytes, and thus the entirety of our data required over 500 gigabytes of storage space. The quantity of data required us to create a system that allowed us to easily sort the files into multiple categories. This can be done retroactively or while the files are being collected.

The first part of building this structure was to have a clear and consistent file naming convention for each video segment. An example of our file naming convention can be found in Figure 1.

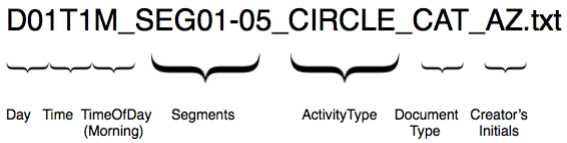

Figure 1: This is an example of a file naming strategy that allows the researcher to provide enough useful information to easily sort a list of files by different category-type (*i.e*. sort by day, time of day, activity type, etc).

We chose these categories for our naming convention based on the types of research questions we were interested in pursuring. We recognize that there are other potential categories to include depending on the project and type of video that is collected. For instance, if a project has a long time-scale, it may be more useful to have a date instead of a day number. A numerical date would work better for the sake of clarity and sorting: *e.g*. 150623 for June 23rd, 2015. The categories we think would be pertinent for a large quantity video data are shown below in Tables 1-4.

Having a file naming convention is immensely useful for navigating the video clips, but even with a clear file name, it is difficult to navigate the files if they are all clumped into one large list. Having the files organized into folders enables easy navigation. There can be as many or as few folders as the research team wants. We recommend titling the folders based on relevant categories in the file name.

Table 1: Days/Sequential Categories Table provides a detailed explanation of each day/sequential-related code in the file name.

Code Example	Description	Purpose
D06	-D for day; followed by a number -If more than 9 days, include "0" before single digits for sorting purposes	-Projects where data collection is on consecutive days and the total collection period is less than a month
W02D6	-W for week; D for day; both are followed by a number	-Classroom data; it allows the researchers to quickly note how far along the class is in the semester
150623	-This is a date -It goes year, month, and then day of month	-Projects spanning months or years

Table 2: Camera Placement and Segments Table provides a detailed explanation of each camera placement/segment-related code in the file name. These descriptions may not be relevant to every type of project. If all of the videos take place in the same setting from one vantage point without segmenting, there would be no reason to include these categories.

Code Example	Description	Purpose
T4	-T for table; followed by a number	-Studying and keeping track of different small groups
SEG05	-SEG for segment; followed by a number	-If a video is broken up into segments for any reason, then it is important to clearly label the segment numbers in the file name.

Table 3: Activity and Video Type Table provides a detailed explanation of each activity/video-related code in the file name. This is not an exhaustive list of all possible video types. We also recommend always including the video type in the file name, even if all of the videos are the same.

Code Example	Description	Purpose
GRPWRK, JRNLG, CIRCL	-GRPWRK for group work, JRNLG for journaling, CIRCL for whole group discussion	-Multiple activities the subject could be working on
INT, GRPINT, HHS	-INT for interview, GRPINT for group interview, HHS for homework help session	-When the subject is only working on one activity

Table 4: Other Categories Table provides a detailed explanation of any additional code in the file name.

Code Example	Description	Purpose
Instructor Code	-This would be an accepted convention for instructors, tutors, etc.	-Tracking specific people
Subject Code	-This would be an accepted convention for the subject	-Tracking a specific aspect of a subject -*e.g.* 2G and 2H for different tutors
Interviewer Code	-This would be an accepted convention for interviewers	-Tracking who conducted the interview

It is possible to choose not to use folders and just store all of the files in one place. For either listing method, creating a spreadsheet can aid organizational structure. If it is the case that creating such a spreadsheet is not worth the man hours to a research team, then the team should merely create a document that describes the labeling system– in fact, this should be done regardless of whether a spreadsheet is created. An example of our spreadsheet is shown below in Table 5.

Table 5: Spreadsheet of Files and Their Categories

Table	Time of day	Segments	Type	Catalog	Field Notes
1	Afternoon	01:05	CIRCLE	D01T1A_SEG01-05_CIRCLE_CAT_AC	Day 1 Metacog Discussion
1	Afternoon	07:23	CIRCLE	D01T1A_SEG07-23_CIRCLE_CAT_AC	Day 1 Metacog Discussion
3	Afternoon	01:24	CIRCLE	D01T3A_SEG01-23_CIRCLE_CAT_RD	Day 1 Metacog Discussion
5	Afternoon	01:23	CIRCLE	See D01T3A_SEG01-23_CIRCLE_CAT_RD	Day 1 Metacog Discussion
0	Morning	01:07	CIRCLE	See D02T3M_SEG01-06_CIRCLE_CAT_NM	Day 2 Morning Field Notes_NM

Labeling and using columns in the spreadsheet allows the research team to sort by the various categories. This lets the team to easily find the types of video clips they are interested in.

Collaborative Aspects

When a project involves the quantity of data described in previous sections, it becomes impossible for all of the work to be done singularly. It is likely that certain parts of the project will be worked on individually, but there is too much to do for any one person. To aid in the collaborative aspect, it is necessary to use an online repository for all *non-video* files. The IMPRESS project chose to organize files via Github, an online repository that allows for

version control. Given the present state of technology, it is not feasible to house all of the files in an online repository; however, it is possible to have them available to stream. An online repository allows the research team to safely store their files and have them available for all members of the collaboration team to work on. This is especially useful for the spreadsheet, as it could be updated on a daily basis during the data collection phase.

The spreadsheet and file naming system provide necessary information for a research team, but more information about each file is always welcome. Cataloging provides both a quick summary of the video as a whole (*e.g*. Video of Class Activity 1), and a short description of what is going on at each time period in the video (*e.g*. 02:40 Corey greets the students, 02:45 the students begin the activity). When members of a research team are working remotely and are all interested in different parts of the data set, having a catalog allows a researcher to quickly figure out if the video has features that he or she is interested in pursuing.

Narrative and Analysis

Our structural and logistical considerations during our data collection phase allowed us to engage with the data and perform analyses quickly and effectively. The analysis our research team engaged utilized a generative writing process. A major part of such a writing process is to write a narrative of a video or set of videos. A narrative aims to tell the story of a part of the data one is interested in. In doing so, it is expected that the researcher is watching the footage with a lens that notices only the actions and scenarios that fall under the frame of interest. The narrative should have sufficient detail so interpretations can be made from the data set, but it should not end up resembling a transcript (though including a short transcript may strengthen the case that the narrative is discussing). We suggest roughly one page of writing for each five minutes of video. This assures that enough of the video's material is recognized without providing excessive information. Upon completion, the narrative should be a detailed, yet brief, overview of a video or set of videos within a frame of interest.

Conclusion

As it becomes easier and cheaper to store data, the need for effective means to organize it grows. We have detailed a system that works for video and audio files. Organization considerations can be easily overlooked in the research process, but our system is simple enough to implement as the data is being collected, and its features enable effective navigation to aid in research analyses.

ACKNOWLEDGEMENTS

This research was made possible by the IMPRESS team at Rochester Institute of Technology. This research was also partially funded by the Research Experience for Undergraduates (REU) at Kansas State University and the Undergraduate Summer Research Program (USRP) at DePaul University.

AUTHOR CONTRIBUTIONS

All authors contributed extensively to this project. N.M. collected the data and wrote the majority of the paper. R.D. created the tables and figure, wrote the Abstract, Narrative and Analysis, and Acknowledgements sections of the paper, and edited the paper throughout various stages of completion.

REFERENCES

Engle, R.A., Conant, F.R., and Greeno, J.C. (2007). "Progressive refinement of hypotheses in video-supported research." In *Video Research in the Learning Sciences*, 239-254. Mahwah, NJ: Lawrence Erlbaum Associates.

Welcome to IMPRESS. (2015). Retrieved April 07, 2016, from https://www.rit.edu/science/impress

Franklin, Scott V. "Internal analysis of 2002 - 2012 RIT student performance and retention" (unpublished).

Mytkowicz, P., Gross, D., & Steinberg, B. (2014). "Assessing Metacognition as a Learning Outcome in a Postsecondary Strategic Learning Course." *Journal of Postsecondary Education & Disability, 27*(1).