iSET: enabling *in situ* & *post hoc* video labeling

Mish Madsen
MIT Media Lab
20 Ames Street
Cambridge, MA 02139
mish@media.mit.edu

Abdelrahman N. Mahmoud
American University in Cairo
AUC Avenue, P.O. Box 74
New Cairo, Egypt 11835
a_nasser@aucegypt.edu

Youssef Kashef
American University in Cairo
AUC Avenue, P.O. Box 74
New Cairo, Egypt 11835
ypx@aucegypt.edu

ABSTRACT
Video annotation is an important component of many behavioral interventions for autistic populations. This demonstration presents the interactive Social-Emotional Toolkit (iSET), a highly portable system for *in situ* video recording and labeling. This tool enables the recording of event labels in a variety of contexts, including behavioral interventions, usability assessments, and interaction studies. With iSET, video can be easily collected and annotated *in situ* with custom labels and reviewed on-site or later, with labels added or removed to assist video analysis. We describe the current usage as a tool enabling a social-behavioral intervention allowing persons with Autism Spectrum Disorders to capture and review expressions of affect during social interactions.

Categories and Subject Descriptors
J.4 [**Computer Applications**]: Social and behavioral sciences – *psychology.*

General Terms
Design, Experimentation, Human Factors, Measurement.

Keywords
Autism, user interaction studies, video annotation, video review.

1. INTRODUCTION
Persons with Autistic Spectrum Disorders (ASD) frequently encounter significant difficulties during social interactions [1]. Behavioral interventions designed to address this issue often rely on the use of acted video depicting unfamiliar actors. Autistic persons[1] may benefit from viewing and discussing naturalistic video in an educational context, but current video-recording systems do not easily facilitate recording and annotation of naturally-occurring social interactions. As a result, video footage of social interactions compiled to assist persons with ASD and similar conditions frequently relies on painstaking manual annotation without a viable real-time alternative.

ASSETS '09, October 26–28, 2009, Pittsburgh, PA, United States.

[1] See Sinclair 1999 for an essay on respectful "autism-first" language at web.syr.edu/~jisincla/person_first.htm through archive.org.

In response to these issues, we have developed a customizable toolkit for i*n situ* and *post hoc* tagging of video. The interactive Social-Emotional Toolkit, iSET, allows caregivers or other observers of an interaction to collect and annotate video *in situ* on an unobtrusive ultramobile computer. Alternatively, iSET allows one or more people to label the same video data *post hoc*, on a portable or desktop platform.

iSET consists of a Samsung ultramobile computer, equipped with an outwards- facing onboard camera running iSET interface, a custom-built application which supports video recording and real-time video annotation (Fig. 1). The iSET interface, written in Visual C++, was developed over several participatory design sessions with persons with ASD and their caregivers in the school setting. The feedback collected was used to create a highly intuitive system for use with and by user with various cognitive and physical deficits.

Figure 1: iSET interface for *in situ* annotation.

The iSET interface is intended as a versatile tool for live and offline video annotation with and by autistic persons; it forms the main real-time component of the iSET intervention, which is designed to support interactive teaching of affective information in a real-world context. The students with ASD who participate in the intervention use the toolkit for a few hours each week of the 15-week intervention to label their interpretation of the expressions of others during social interactions by selecting different labels for real-time video recordings, such as "happy" or "confused."

Multiple engaging games have been developed to focus the participants' attention during this time; for instance, students might be asked to find several different people looking "unsure", or record the same person expressing "interested" in different contexts. The video is then reviewed independently, with a caregiver, or with the social partner from the original interaction.

2. *IN SITU* VIDEO ANNOTATION

During live use, the iSET interface (Fig. 2) may be configured to represent any number of custom labels. The recording and annotation process takes place by allowing the user to press colorful labeling buttons (Fig. 2) in order to label the video clips during the recording session; labels can be applied at any point during live recording. Following each recording session, the interface produces a comma-separated log file containing the recorded labels and the corresponding times.

The labels can be customized via a control panel before or while the recording takes place. For instance, in our intervention for persons with ASD, we have customized these labels to indicate Ekman's "basic" emotional states [2], such as "happy", "sad", and "surprised", as well as more complex cognitive states suggested by Baron-Cohen's Mind Reading DVD [3], such as "agreeing" and "concentrating."

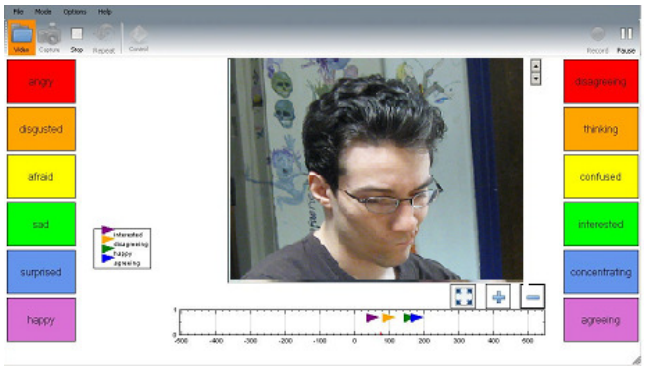

Figure 2: iSET interface's *post hoc* labeling window, including timeline markers and video navigation options.

3. *POST HOC* VIDEO ANNOTATION

Once the original video is recorded, iSET interface allows the recorder of the video to review it on the handheld device or on another computer in order to revise and improve the labels assigned to video clips (Fig. 2). The events that were originally recorded appear in a timeline below the video, and a legend is provided to indicate which flag color on the timeline corresponds to each labeled expression. Clicking on the timeline allows the user to navigate through the video or review specific sequences.

With the *post hoc* review system, a user may open a video and choose to open the original timeline of events or create a new one. For instance, two study participants might evaluate the same video in order to provide their own input on how facial expressions might indicate different combinations of the available emotions. The video could then be watched alongside the labelers by the recorded participant, allowing that participant to gain a greater understanding of how his/her facial expressions are perceived by those around him, and allowing the labelers and recorded participant to engage in dialogue about the function of facial expression during a social interaction.

In the case of the iSET autism intervention, there was a need to establish 'ground truth' of the emotional content recorded from social interactions in order to form an effective, naturalistic corpus of videos for the intervention evaluation metric. Video labelers were recruited online and shown specific videos collected in the classrooms using the iSET toolkit [4]. (Each participant in the iSET intervention gave explicit consent/assent for this use of their video, approved by MIT's IRB.) For each clip, viewers were presented with a set of labels representing possible interpretations, and videos with high levels of agreement between labelers were used to create the evaluation metric.

Both student and caregiver users of iSET were enthusiastic about collecting video and labeling it through the iSET system. More than 4000 short segments expressing single emotions have been collected so far through use of the iSET system, segmented automatically from over 15 hours of video. These videos feature autistic student participants as well as their teachers. The final evaluation metric will contain two clips demonstrating each of 8 emotions (as indicated by VidL-based consensus) for each of the 21 participants. During pre- and post-testing, participants will attempt to identify the emotions expressed in clips of themselves; each participant will also view clips of each emotion featuring of two teachers and two peers, and of the emotions demonstrated by an unfamiliar adult.

4. OTHER USES OF THE ISET TOOLKIT

The iSET toolkit may be applicable to other interventions where painstaking manual annotation is currently in use. For instance, it may be useful in a user-device interaction study taking place over an extended period of time to record certain types of user emotion, e.g. "frustration" or "excitement", or user interaction with certain functionalities, e.g. "text to speech activated" or "camera accidentally turned off." This approach will allow both easy 'replay' of interesting events and simple collection of annotation data for later quantitative analysis.

5. CONCLUSION

The iSET project has developed an important new tool to make it easier and more efficient to annotate video gathered in real-world settings. The iSET interface makes it simple to label videos in real-time and review labels in the original recording context.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 055541. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors thank the participants and staff at the Groden Center for participating.

7. REFERENCES

[1] National Institute of Neurological Disorders and Stroke. (2007c). *Autism Spectrum Disorders.*

[2] Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion, 6* (3-4): 169-200.

[3] Baron-Cohen, S., et al. (2004). Mind Reading: The Interactive Guide to Emotions. Jessica Kingsley Publishers.

[4] Eckhardt, M. and Picard, R. *A More Effective Way To Label Affective Expressions*. In Proceedings of Affective Computing and Intelligent Interaction ACII-2009. Amsterdam, 2009.