Crowdsourcing Micro-Level Multimedia Annotations: The Challenges of Evaluation and Interface

Sunghyun Park[1], Gelareh Mohammadi[2], Ron Artstein[1], and Louis-Philippe Morency[1]

[1]Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Playa Vista, CA 90094, United States
{park, artstein, morency}@ict.usc.edu

[2]Idiap Research Institute
CP 592 – 1920 Martigny, Switzerland
École Polytechnique Fédérale de Lausanne - EPFL
1015-Lausanne, Switzerland
gelareh.mohammadi@epfl.ch

ABSTRACT

This paper presents a new evaluation procedure and tool for crowdsourcing micro-level multimedia annotations and shows that such annotations can achieve a quality comparable to that of expert annotations. We propose a new evaluation procedure, called MM-Eval (Micro-level Multimedia Evaluation), which compares fine time-aligned annotations using Krippendorff's alpha metric and introduce two new metrics to evaluate the types of disagreement between coders. We also introduce OCTAB (Online Crowdsourcing Tool for Annotations of Behaviors), a web-based annotation tool that allows precise and convenient multimedia behavior annotations, directly from Amazon Mechanical Turk interface. With an experiment using the above tool and evaluation procedure, we show that a majority vote among annotations from 3 crowdsource workers leads to a quality comparable to that of local expert annotations.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *performance evaluation (efficiency and effectiveness).*
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *graphical user interface, evaluation/methodology.*
H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – *computer-supported cooperative work, evaluation/methodology, web-based interaction.*

General Terms

Design, Experimentation, Human Factors, Measurement.

Keywords

crowdsourcing, Amazon Mechanical Turk, video annotation, behavior annotation, inter-coder agreement, inter-rater reliability, OCTAB.

1. INTRODUCTION

Annotating multimedia content is becoming an important part of many recent research problems, including multimedia event recognition [17], video retrieval and classification [11], and human behavior analysis [16]. Supervised learning approaches applied to these research problems usually require a large number of annotated video sequences. While some of these algorithms are applied at the video or scene level (referred to as macro-level annotations), many of these problems need micro-level annotations, where the precise start and end of an event or a behavior needs to be annotated. These annotation efforts can be extremely time-consuming and require a significant budget.

CrowdMM '12, October 29, 2012, Nara, Japan.

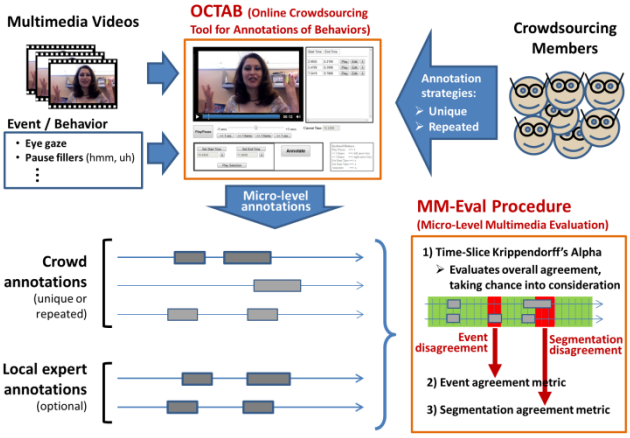

Figure 1. Overview of our approach for crowdsourcing micro-level video annotations, with a focus on our new interface, called OCTAB, and evaluation procedure for multimodal behavior annotations, called MM-Eval.

In recent years, there has been an explosive growth in the research and use of crowdsourcing, fueled by convenient online crowdsourcing environments like Amazon Mechanical Turk. In the research community, crowdsourcing is already being actively used for many types of tasks, including image labeling [15] and linguistic annotations [14]. When using crowdsourcing for micro-level multimedia annotations, two main challenges emerge: evaluation and interface. While many evaluation metrics have already been proposed to assess the quality and agreement of macro-level annotations, there has been limited work to evaluate the agreement in micro-level multimedia annotations. There is also a need of web interfaces that allow crowd workers to accurately and efficiently annotate micro-level events and behaviors while keeping the interfaces simple and intuitive.

In this paper, we propose a new evaluation procedure, called MM-Eval (Micro-level Multimedia Evaluation), which applies the widely used Krippendorff's alpha metric [8] to micro-level multimedia annotations and introduce two new metrics to evaluate the types of disagreement between coders (see Figure 1). We also introduce OCTAB (Online Crowdsourcing Tool for Annotations of Behaviors), a web-based annotation tool that allows precise and convenient multimedia behavior annotations, directly from Amazon Mechanical Turk interface. Finally, we present an

extensive set of experiments evaluating the feasibility of crowdsourcing micro-level multimedia annotations and show that a majority vote with 3 crowdsourced repeated annotations leads to a higher agreement with individual expert annotators than the agreement between the experts themselves.

2. RELATED WORK

Crowdsourcing has gained much attention lately, and a survey paper by Yuen et al. [27] and another by Quinn and Bederson [18] present a general overview of the topics on crowdsourcing and human computation. Regarding Amazon Mechanical Turk, Mason and Suri [13] provided detailed explanations on using the platform for conducting behavioral research, and Ross et al. [21] showed changing demographics of the people using the platform.

Quality control is a critical issue with crowdsourcing. Downs et al. [4] and Rashtchian et al. [19] showed the benefit of a screening/qualification process, Le et al. [10] showed an approach of adding a training period in designing a study, and Sheng et al. [22] explored repeated labeling of data for more reliability. By comparing annotations (none of them on videos) obtained with crowdsourcing and those with expert annotators, several [5, 6, 12, 15, 19, 23] have reported across different domains that they could obtain good quality annotations through crowdsourcing. In our work, we incorporate most of these quality control measures and further show experimental results of comparing micro-level annotations of videos obtained with crowdsourcing with those done by expert annotators.

As for crowdsourcing video-related tasks, Wu et al. [26] worked on obtaining video summarizations, Biel and Gatica-Perez [2] on macro-labeling impressions of vloggers in videos, and Riek et al. [20] on macro-labeling social contexts in video scenes, but none of them were concerned with micro-level annotations. Probably most relevant pieces of work in terms of our web interface were done by Vondrick et al. [25] and Spiro et al. [24], whose interfaces allowed micro-level video annotations and were also used with Amazon Mechanical Turk. However, their interfaces put an emphasis on motion tracking, while our interface is more concerned with annotating behavioral events in videos. Although there are quite a number of software for making annotations on videos [3], such full-fledged tools are not suitable to be used for crowdsourcing due to a relatively steep learning curve and the difficulty in incorporating it into crowdsourcing platforms like Amazon Mechanical Turk.

Krippendorff's alpha has been previously used to measure inter-coder reliability of video annotations both at a macro-level [20] (label on the whole video clip) and micro-level [7] . In this paper, we follow the approach taken in [7] at a micro-level, but we further explore the stability of the alpha at different temporal resolutions. We also propose two new measures to supplement the alpha because it cannot show the types of disagreement between coders.

3. MM-Eval: Evaluation Procedure for Micro-Level Multimedia Annotations

In this section, we present our guidelines to evaluate micro-level video annotations of human behaviors or multimedia events. For the purpose of this paper, we suppose that we have expert annotations for each video and our purpose is to evaluate the validity and agreement of the annotations coming from non-experts via crowdsourcing. However, the procedure and metrics described herein could also be used for local annotations and extended for scenarios where no expert annotations are available. These research questions are part of our future work.

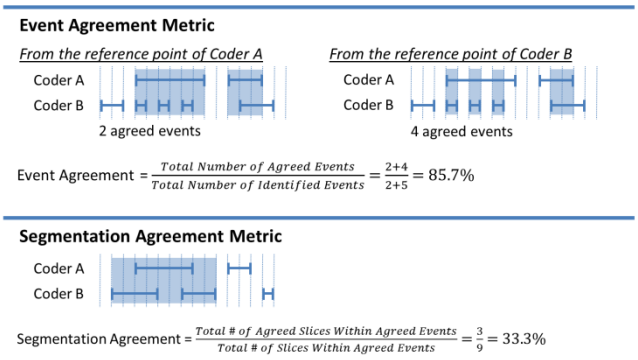

Figure 2. Definition of the event and segmentation agreement metrics with examples.

Figure 1 shows an overview of our approach where crowdsourcing workers are asked to annotate a specific behavior or event in a series of videos. The goal of our MM-Eval (Micro-level Multimedia Evaluation) procedure is to quantify the quality and agreement of these crowdsourced annotations and also give guidelines for future annotations as to how many coders per video are required. For example, if the agreement of unique crowdsourced annotations with expert annotations is shown to be high, then one crowd coder per video and behavior/event will be sufficient in the future. The following subsections explain the three main steps of MM-Eval, including the building of crowdsourced annotations (unique or majority), the overall agreement analysis by applying Krippendorff's alpha to multimedia annotations, and the analysis of coder disagreement types using two new metrics.

3.1 Crowd Annotation

As a pre-processing step to the annotation evaluation *per se*, we assume repeated annotations are obtained from crowdsourcing workers per video and create combined annotations from the crowdsourced annotations by simple majority voting. By testing different levels of majority voting, we can analyze the advantage of having multiple annotators per video. In this paper, we focus on two approaches: unique and majority.

Unique This first approach keeps all original crowdsourced annotation sets separately (three separate sets if three workers annotated per video) and applies the evaluation metrics. This approach can be seen as a case of the majority approach, when the number of set is 1. We use this approach to test if one set of crowdsourced annotations is sufficient.

Majority This second approach computes a combined crowdsourced annotations using majority voting. If three sets of crowdsourced annotations are available (as is the case in our experiments), as long as two sets agree that there should be an annotation, then the majority voting set will include this annotation.

3.2 Time-Slice Krippendorff's Alpha

Our first measure is Krippendorff's alpha [8], which is a generalized chance-corrected agreement coefficient that can be calculated between two or more annotators. The general formula for the alpha is the following:

$$\alpha = 1 - \frac{D_o}{D_e} \quad (1)$$

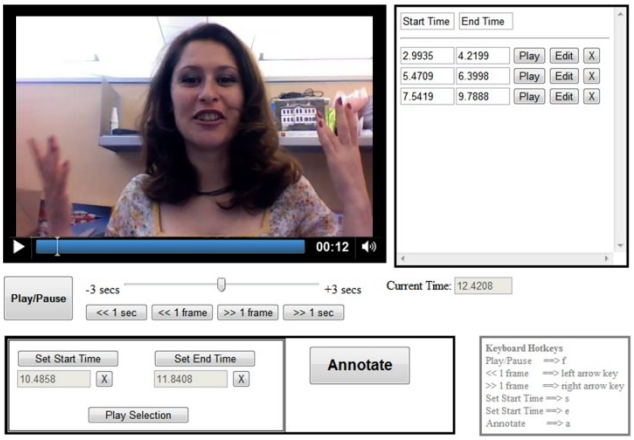

Figure 3. OCTAB (Online Crowdsourcing Tool for Annotations of Behaviors).

where D_o, or observed disagreement, is the amount of pairwise disagreement observed between the annotators, and D_e, or expected disagreement, is the level of disagreement expected by chance as calculated from the data. The coefficient alpha itself is a measure of agreement ranging from -1 to 1, where 1 is perfect agreement (zero observed disagreement), 0 is chance-level agreement, and values lower than 0 indicate systematic disagreement.

The alpha works by looking separately at the agreement on individual annotation instances. For micro-level annotations, we treat each time slice (e.g., 1 frame per slice) as a separate annotation instance, with a binary annotation indicating presence or absence of a specific behavior (such as a frown). While it is the case that adjacent frames tend to have similar annotations, our experiments (in Section 6) show that the alpha is not very sensitive to the sampling rate. The agreement is calculated separately for each annotated behavior.

Applying the alpha to individual time slices means that the measure can only assess whether the annotators agree that at a certain time point a behavior takes place, not whether they agree about the segmentation or the individuation of behaviors (whether a certain time span contains one or two instances of a frown); this drawback has been pointed out by Krippendorff [9]. To supplement the alpha, we devised two additional measures which are intended to capture agreement on the individuation of annotated behaviors.

3.3 Disagreement Type Analysis

As mentioned in the previous section, the Time-Slice Krippendorff's alpha does not differentiate between a disagreement caused by misalignment of the annotations or direct event disagreement (see Figure 1). To better understand these annotation differences, we propose two new metrics:

Event Agreement Metric An agreed event is defined as when there is an overlap of identified events in two annotations. In other words, agreed events are those that both annotators jointly identified. Depending on which annotation is taken as the reference point however, the number of agreed events could be different (see Figure 2). For this reason, we compute the percentage of agreed behavior events between the two annotations by dividing the total number of agreed events from both reference points by the total number of identified events from both reference points.

Segmentation Agreement Metric Another informative measure in gauging the agreement between two annotators is to see how precisely they segmented the boundaries of the same annotation events. To compute the segmentation precision, we look at the time windows of agreed behavior events from both reference points combined and compute agreement within the time windows only (see Figure 2). The percentage is computed by dividing the number of agreed time slices by the number of total time slices within the time window.

4. OCTAB: WEB-BASED VIDEO ANNOTATION TOOL

We developed OCTAB [1] (Online Crowdsourcing Tool for Annotations of Behaviors), which is a web-based interface that allows an annotator to conveniently navigate in a video to annotate micro-level events or human behaviors (see Figure 3). OCTAB is intended for annotating a single behavior on a single video at a time, and it is based on JavaScript and provides all the basic functionalities of a web video player. We considered the following three main aspects in our design of OCTAB:

Precision For accurate micro-level annotations on videos, annotators need to have frame-level precision in identifying the start and end time of an event. To address this requirement, the interface provides the annotator with 4 buttons for moving 1 second backward/forward and 1 frame backward/forward from the current time in the video, as well as a slider bar that offers frame-level navigation in the range from -3 to +3 seconds. Once the annotator identifies an event of a behavior to annotate, he/she can use the navigation control buttons to pinpoint and select the behavior event's start and end time. Then, he/she can play the selection to verify and press a button to save the selection as an annotation.

Integrability (with Amazon Mechanical Turk) Popular annotation software applications like ELAN or ANVIL [3] allow annotators to make sophisticated annotations on video and audio files, but they are not suitable for the purpose of crowdsourcing. They have a relatively steep learning curve to use and cannot be used with online crowdsourcing platforms like Amazon Mechanical Turk. OCTAB was written directly in HTML so that it can be easily used to create a template task page in Mechanical Turk.

Usability Annotating videos often involves moving around in a video to check, re-evaluate and edit previously made annotations (especially in the beginning). A special section in OCTAB displays a list of all saved annotations, and the annotator can always go back and work on previously made annotations by replaying, editing or deleting any annotations. For convenience and speed in making annotations, most controls in the interface have hotkeys associated with them, and the interface's functionalities are kept to the minimal level with an intuitive layout to minimize confusion.

5. EXPERIMENT

In our experimental design, we put our focus on testing the potential of crowdsourcing for obtaining micro-level annotations of various human behaviors displayed in videos. Agreement between two in-house expert annotators was taken as the standard level of quality, and this experiment studied how comparable the

[1] Will be made freely available for research at http://multicomp.ict.usc.edu

crowdsourced annotations could be to those made by expert annotators using our OCTAB interface.

5.1 Multimedia Videos

From YouTube, which is a video-sharing website where users can upload and share videos, about 360 videos of people giving movie reviews were collected and each was given a sentiment score by expert annotators. From those videos, 20 videos were selected for this study that were short and both gender-balanced and sentiment-balanced (to have various expressions), and 3 videos were randomly selected to be used for task instructions. Each video showed a frontal, upper-body shot of a different person talking. Since all of the videos appeared to have been recorded using a webcam, the overall quality of the videos were not ideal but still fair enough to discern various facial expressions and eye gaze. For the 20 videos that were used in the actual experiments, the frame rate was at 25 frames per second and the video length ranged from 60 to 180 seconds, averaging at 138 seconds.

5.2 Annotated Behaviors

From behaviors that were relatively common and frequent in all the videos, we selected 4 different types of behaviors to annotate based on their variety (one for the eye, one for the facial expression, one for the head movement, and one for verbal), and they are the ones that are also frequently annotated for research involving human behavior analysis. The descriptions of the behaviors below were adapted from the MUMIN multimodal coding scheme [1].

- Gaze Away: eye gaze is directed away from the camera.
- "Um" and "Uh" (pause fillers): the person says "um..." or "uh..."
- Frown: the eyebrows contract and move toward the nose.
- Headshake: a repeated rotation of the head from one side to the other.

5.3 Experimental Design

Amazon Mechanical Turk (AMT) is arguably the most well-known and widely used environment for crowdsourcing. The main idea behind AMT is to distribute small tasks at which humans are good and computers are still bad to a crowd of "workers" worldwide. Using AMT's web interface, the "requesters" can design and publish tasks online, which are called HITs for Human Intelligence Tasks. In designing a HIT, the requesters can set various options to restrict access to specific kinds of workers, set the number of unique workers to work on it, and set the amount of monetary reward. Once HITs are published online, the workers can work on them and submit the results, which the requesters can either approve or reject. In addition, the requesters can grant qualifications to particular workers and design HITs to be accessible only to those workers with certain qualifications. The reader is referred to [13] for more detail on using AMT.

5.3.1 Task Design

When designing HITs, a HIT template can be created, and one can define variables whose values will vary from HIT to HIT, which becomes very useful in creating a batch of similar HITs. First, we created a HIT template with OCTAB integrated. In the HIT template, we used a variable for the file address of the video to play. Then, the template was given an input file specifying the links to the 20 videos to be used in place of the variable. The videos were hosted elsewhere with public access.

Using the template, a batch of 60 HITs were created for each behavior of gaze away, "um" and "uh", frown, and headshake. We expected that annotating the gaze away behavior would be significantly more time-consuming than the rest, and used 1-minute cropped versions of the videos for the behavior. Furthermore, we allowed each HIT to be completed by 3 unique workers. Since we had 20 videos to annotate per behavior and had 3 unique workers annotate each video per behavior, 60 HITs were created per behavior. The monetary reward was based on an effective hourly wage of $4. In total, 240 HITs were created with $300 for an estimated 70 hours of work.

The instruction defining each type of behaviors was kept to the minimal level. Along with the simple word description outlined in Section 5.2, only 3 video snippets (different videos from the 20 videos used for the actual experiments) of positive behavior events were shown for each behavior.

5.3.2 Worker Recruitment and Qualification

Through a posting on mTurk forum [2], which is an online community of active AMT users, 16 workers were invited to try a 30-minute tutorial HIT for $2. The tutorial HIT involved reading the instructions for behavior annotation task in general and becoming familiar with OCTAB interface by annotating a 1-minute video for gaze away behavior. This qualification step can be seen as a minimal practice session because our main purpose lay in ensuring that the workers understood annotation tasks in general and that they familiarized themselves with using OCTAB interface. Out of the 16 workers who completed the tutorial HIT, 14 workers (5 based in the U.S. and 9 in India) were granted a qualification to access the real HITs, and 2 workers were screened out because they bordered on spammers. For the actual experiment, all of the 240 HITs completed by 12 qualified workers (2 workers did not do any HITs before the experiments were completed) were approved without rejections. In other words, the initial qualification process was the only quality assurance measure taken in this study. However, in return for a relatively favorable effective hourly wage of $4 (based on a quick poll on the qualified workers) and a promise of no rejections, the workers also promised to be sincere in their work.

5.4 In-house Expert Annotators

For comparison, we recruited two expert in-house annotators. Each annotator went through the same process of completing the tutorial HIT and used the same environment on AMT for the actual annotation tasks. For each annotator, the same batch of 20 HITs were created for each behavior (total 80 HITs) as the ones created for the workers on AMT. The annotators were not given any more detailed description of the tasks, and they did not communicate about the annotation tasks with each other.

5.5 Methodology

In these experiments, we compared the performance of three annotation approaches: experts, unique, and majority.

Experts We had two local expert annotators (not from crowdsourcing) who each produced a complete set of annotations (20 videos, 4 behaviors). The agreement between these two local expert sets is considered as the gold standard in our experiments. We refer to these sets as *expert*s in the result section.

Crowdsourced Unique From the crowdsourcing workers on AMT, we obtained 3 annotation sets from different workers per

[2] http://www.mturkforum.com

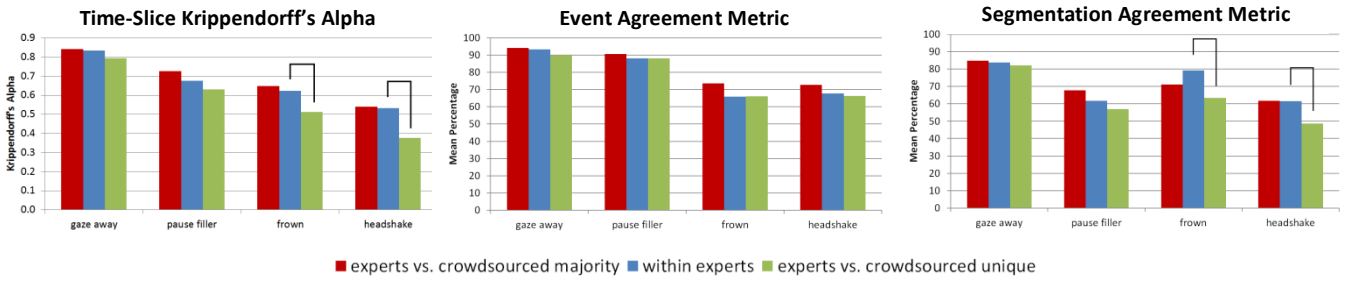

Figure 4. Agreement between *experts* and *crowdsourced majority*, within *experts*, and between *experts* and *crowdsourced unique* (brackets highlight places with strong differences).

video per behavior. By randomly permuting the order in the 3 annotation sets, we created 3 complete sets of crowdsourced annotations, which we refer to as *crowdsourced unique*.

Crowdsourced Majority As described in Section 3.1, the 3 crowdsourced annotation sets can be combined to make another complete set using majority voting where a sample is judged annotated if at least 2 out of 3 annotators agreed. We refer to this set as *crowdsourced majority*.

We compared the annotation agreement in three different combinations: (1) within *experts* so that we have a baseline, (2) *crowdsourced unique* vs. *experts* to know if one crowdsourcing worker is sufficient, and (3) *crowdsourced majority* vs. *experts* to know if the majority over three workers is sufficient. For all three combinations, we calculated the Time-slice Krippendorff's alpha, the event agreement metric, and the segmentation agreement metric.

6. RESULTS AND DISCUSSIONS

This section highlights five main research problems studied during our experiments: sensitivity analysis of Time-Slice Krippendorff's alpha measure to test its stability, the performance of *crowdsourced unique* and *crowdsourced majority* compared to *experts* as part of our MM-Eval procedure, the analysis of the types of disagreement using our two new metrics, and the usability of OCTAB interface.

Time-Slice Krippendorff's Alpha For all behaviors, the Time-Slice Krippendorff's alpha was shown to be a stable measure that stayed consistent across different sizes of the time slices, and we show the results for gaze away and frown behaviors in Figure 5. For this experiment, annotation sets created at a slower frame rate were up-sampled using a majority voting approach, where each time slice was considered to be annotated if at least 50% of the slice was annotated.

Performance of *Crowdsourced Majority* Figure 4 (left) shows comparable (if not higher) agreement for *crowdsourced majority* vs. *experts* to the agreement for *within experts* for all 4 behaviors. This suggests that we can obtain annotations that are of comparable quality to expert annotations if we obtain at least 3 repeated annotations by the workers and perform majority voting.

Performance of *Crowdsourced Unique* For the annotation sets obtained with 1 worker per video (*crowdsourced unique*), their agreement with the expert annotation sets was comparable to within *experts* for gaze away and "um" and "uh" (pause filler) behaviors. This is a really exciting result since it means that these behaviors can be coded with only one set of crowdsourced annotations. For the frowning and headshake annotations, we observed a lower agreement between *crowdsourced unique* and

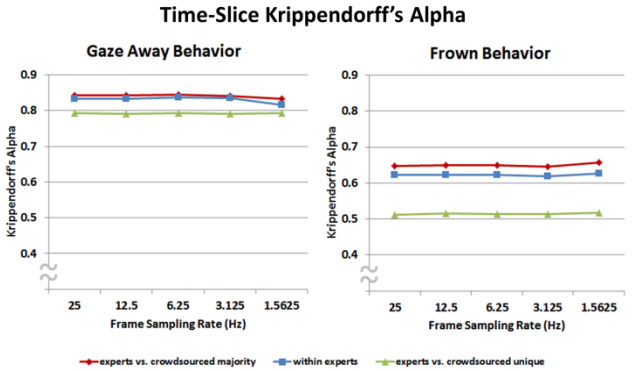

Figure 5. Sensitivity Analysis of Time-Slice Krippendorff's alpha across different frame sampling rates.

experts. To better understand the source of these disagreement, we looked at the new disagreement metrics introduced in Section 3.3.

Disagreement Analysis Figure 4 (middle) shows that, for all behaviors, even without training, the crowdsourcing workers were almost no different than expert annotators in recognizing and identifying the four behavioral events in the videos. However, Figure 4 (right) shows that the segmentation agreement between the workers and the experts was lower for frown and headshake behaviors. This is a really interesting results which gives us insight on how to improve performance of future crowdsourced annotations. The tutorial should better emphasize on how to properly annotate the start and end time of these behaviors to improve segmentation performance. By doing so, we may be able to boost the agreement and possibly have only one crowdsourcing worker annotate per video for frown and headshake behaviors as well and still achieve a high quality.

OCTAB's Usability The 12 workers who completed the HITs were given a short survey on the overall experience. On a 7-point Likert scale to rate how easy it was to understand OCTAB interface (from very difficult at 1 to very easy at 7) and its convenience (from very inconvenient at 1 to very convenient at 7), the mean scores were 6.25 and 6.42, respectively. As can be expected from the agreement results, the workers reported that frown and headshake behaviors were more difficult to annotate than gaze away and "um" and "uh" behaviors.

7. CONCLUSIONS

In this paper, we introduced OCTAB interface that can be used with Amazon Mechanical Turk to crowdsource micro-level multimedia annotations, and we proposed a new evaluation procedure, MM-Eval, to evaluate these annotations. Our

experiments, as part of the MM-Eval procedure, show that the Time-Slice Krippendorff's alpha is not sensitive to the sampling rate. Moreover, for annotating various human behaviors in videos, our results show that a majority vote among annotations from 3 crowdsource workers leads to a quality comparable to that of local expert annotations. Lastly, for relatively unambiguous behaviors like gaze away and pause fillers, our results suggest that having one crowdsourcing worker annotate per video (non-repeated annotation) is enough to obtain annotations of comparable quality to those by expert annotators.

8. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1118018 and the U.S. Army Research, Development, and Engineering Command. 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 or the Government.

9. REFERENCES

[1] Allwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navarretta, C., and Paggio, P. 2004. The MUMIN multimodal coding scheme. *NorFA yearbook 2005*, 129-157.

[2] Biel, J. I. and Gatica-Perez, D. 2012. The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers. In *Proc. of the 6th Int'l AAAI Conf. on Weblogs and Social Media*. ICWSM '12.

[3] Dasiopoulou, S., Giannakidou, E., Litos, G., Malasioti, P., and Kompatsiaris, Y. 2011. A survey of semantic image and video annotation tools. In *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, G. Paliouras, C. Spyropoulos, and G. Tsatsaronis, Ed. Lecture Notes in Computer Science. Springer Berlin, 196-239.

[4] Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. 2010. Are your participants gaming the system?: Screening mechanical turk workers. *In Proc. of the 28th Int'l Conf. on Human Factors in Comp. Sys.* CHI '10, 2399-2402.

[5] Gao, Q. and Vogel, S. 2010. Consensus versus expertise: A case study of word alignment with Mechanical Turk. In *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Lang. Data with Amazon's Mech. Turk*. CSLDAMT '10, 30-34.

[6] Hsueh, P. Y., Melville, P., and Sindhwani, V. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proc. of the NAACL HLT 2009 Workshop on Active Learning for Natural Lang. Processing*, 27-35.

[7] Kang, S., Gratch, J., Sidner, C., Artstein, R., Huang, L., Morency, L. 2012. Towards building a virtual counselor: Modeling nonverbal behavior during intimate self-disclosure. In *Proc. of the 11th Int'l Conf. on Autonomous Agents and Multiagent Systems*. AAMAS '12, 63-70.

[8] Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage, Beverly Hills, CA.

[9] Krippendorff, K. 1995. On the reliability of unitizing contiguous data. *Sociological Methodology*, 25, 47-76.

[10] Le, J., Edmonds, A., Hester, V., and Biewald, L. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *Proc. of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Eval.* CSE '10, 21-26.

[11] Lew, M. S., Sebe, N., Djeraba, C., and Jain, R. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM T. Multim. Comput.* 2, 1 (Feb. 2006), 1-19.

[12] Marge, M., Banerjee, S., and Rudnicky, A. I. 2010. Using the Amazon Mechanical Turk for transcription of spoken language. In *Proc. of the 2010 IEEE Int'l Conf. on Acoustics Speech and Sig. Process.* ICASSP '10, 5270-5273.

[13] Mason, W. and Suri, S. 2011. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (2012), 1-23.

[14] Novotney, S. and Callison-Burch, C. 2010. Cheap, fast and good enough: automatic speech recognition with non-expert transcription. In *Proc. of the 2010 NAACL Human Lang. Tech.* HLT '10, 207-215.

[15] Nowak, S. and Ruger, S. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proc. of the Int'l Conf. on Multimedia Info. Retrieval*. MIR '10, 557-566.

[16] Pantic, M., Pentland, A., Nijholt, A., and Huang, T. 2006. Human computing and machine understanding of human behavior: A survey. In *Proc. of the 8th Int'l Conf. on Multimodal Interfaces*. ICMI '06, 239-248.

[17] Poppe, R. 2010. A survey on vision-based human action recognition. *Image Vision Comput.* 28, 6 (Jun. 2010), 976-990.

[18] Quinn, A. J. and Bederson, B. B. 2011. Human computation: A survey and taxonomy of a growing field. In *Proc. of the 2011 Annual Conf. on Human Factors in Comp. Sys*. CHI '11, 1403-1412.

[19] Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Lang. Data with Amazon's Mech. Turk*. CSLDAMT '10, 139-147.

[20] Riek, L., O'Connor, M., and Robinson, P. 2011. Guess what? A game for affective annotation of video using crowd sourcing. In *Proc. of the 4th Int'l Conf. on Affective Computing and Intelligent Interaction*. ACII '11, 277-285.

[21] Ross, J., Irani, L., Silberman, M., Zaldivar, A., and Tomlinson, B. 2010. Who are the crowdworkers?: Shifting demographics in Mechanical Turk. In *Proc. of the 28th Int'l Conf. Ext. Abst. on Human Factors in Comp. Sys*. CHI EA '10, 2863-2872.

[22] Sheng, V. S., Provost, F., and Ipeirotis, P. G. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*. KDD '08, 614-622.

[23] Snow, R., O'Connor, B., Jurafsky, D. and Ng, A. Y. 2008. Cheap and fast but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proc. of the Conf. on Empirical Methods in Natural Lang. Proc.* EMNLP '08, 254-263.

[24] Spiro, I., Taylor, G., Williams, G., and Bregler, C. 2010. Hands by hand: Crowd-sourced motion tracking for gesture annotation. In *Proc. of the 2010 IEEE Computer Vision and Pattern Recog. Workshops*. CVPRW '10, 17-24.

[25] Vondrick, C., Ramanan, D., and Patterson, D. 2010. Efficiently scaling up video annotation with crowdsourced marketplaces. *Computer Vision - ECCV 2010* 6314, 610-623.

[26] Wu, S. Y., Thawonmas, R., and Chen, K. T. 2011. Video summarization via crowdsourcing. In *Proc. of the 2011 Annual Conf. Ext. Abst. on Human Factors in Comp. Sys.* CHI EA '11, 1531-1536.

[27] Yuen, M. C., King, I., and Leung, K. S. 2011. A survey of crowdsourcing systems. In *Proc. of the IEEE 3rd Int'l Conf. on Social Computing*. SocialCom '11, 766-773.