Glance: Enabling Rapid Interactions with Data Using the Crowd

Walter S. Lasecki[1], **Mitchell Gordon**[1], **Steven P. Dow**[2], **and Jeffrey P. Bigham**[2]

ROC HCI, Computer Science[1]
University of Rochester
{wlasecki, mgord12}@cs.rochester.edu

Human-Computer Interaction Institute[2]
Carnegie Mellon University
{spdow, jbigham}@cs.cmu.edu

ABSTRACT

Behavioral coding is a common technique in the social sciences and human computer interaction for extracting meaning from video data. Since computer vision cannot yet reliably interpret human actions and emotions, video coding remains a time-consuming manual process done by a small team of researchers. We present Glance, a tool that allows researchers to rapidly analyze video datasets for behavioral events that are difficult to detect automatically. Glance uses the crowd to interpret natural language queries, then aggregates and summarizes the content of the video. We show that Glance can accurately code events in video in a fraction of the time it would take a single person. We also investigate speed improvements made possible by recruiting large crowds, showing that Glance is able to code 80% of an hour-long video in just 5 minutes. Our demo will allow participants to define their own events occurring in videos and get feedback within an about 10 to 20 seconds. Rapid coding allows participants to have a "conversation with their data" to rapidly develop and refine research hypotheses in ways not previously possible.

Author Keywords

data analysis, subjective coding, crowdsourcing, video

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation: Misc.

INTRODUCTION

Behavioral coding of video is a common method for researchers in the social sciences to gain insight into interactions [3]. An HCI researcher might spend days or weeks coding videos of user interactions as a first step toward developing a theory to help explain those interactions. Generally, behavioral coding is difficult for computer vision to label because it requires an understanding of human behavior and context. The standard practice is for people to manually code video, which usually takes at least 2-3x the play time of the video itself, to identify occurrences of just a single event [4].

Currently, researchers form hypotheses early and are then locked in, since the turnaround time for coding these events is measured in days. This length also makes it costly (in terms of time and effort, which often leads to the need for hiring another person to complete the task). In contrast, our approach lets researchers easily code their own video interactively, allowing them to revise or revisit questions they ask of their data. By parallelizing the process across multiple crowd workers, the coding process can be reduced down to a small fraction (˜10%) of video playing time. Additionally, as the video is processed, researchers get samples of the video coding done within seconds, allowing them to monitor progress, and decide when and how to proceed. We evaluated our system by coding several types of behavioral events in video from the 2012 Presidential debates.

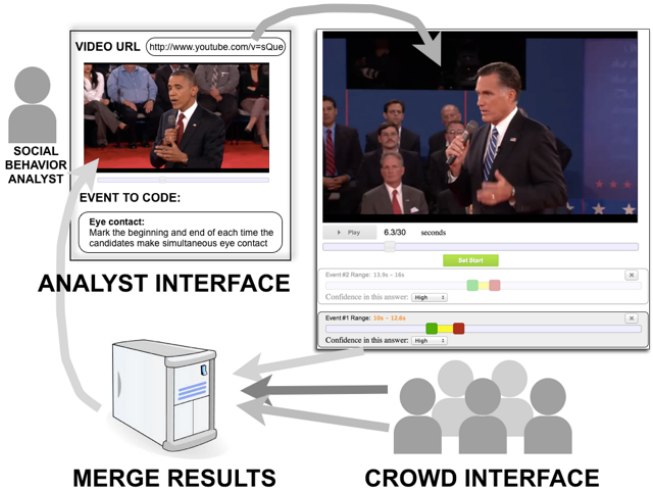

Figure 1. Glance codes behavioral events in video quickly and accurately. When a question is asked, small clips from the video are sent to the crowd workers who label events in parallel. Their answers are then merged together and forwarded back quickly.

Experiments have demonstrated that Glance is able to accurately identify an average of over 99% of events from video, and can mark the duration of these events with less than a 1 second margin of error in terms of when they occur and for how long, within 25% of the time it takes to play the video.

Pushing behavioral analysis of video data to near interactive speeds holds the promise of dramatically expanding the kinds of interaction that analysts can have with their data. Glance can visualize results in real-time, allowing analysts to issue a new query, or reframe their original one after seeing initial results. The resulting conversational style of interaction allows analysts to more quickly explore, develop and refine research hypotheses in ways that are not feasible today.

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.
ACM 978-1-4503-2474-8/14/04.
http://dx.doi.org/10.1145/2559206.2574817

Our contributions can be summarized as follows:

- We present a working, demonstrable system, Glance, that leverages the crowd for rapid behavioral coding of videos.
- We provide experimental evidence via a large study with Amazon Mechanical Turk workers that the crowd can quickly and accurately identify events in video.
- We introduce the idea of using the crowd to enable an analyst to have a "conversation with their data", and demonstrate that the turn-around speeds necessary to support it is possible with the available crowd on Mechanical Turk.

RELATED WORK

Glance uses crowdsourcing to allow users to rapidly iterate on a video coding task. As such, it draws from prior work in both behavioral coding and crowdsourcing.

Behavioral Observation and Coding

Behavioral coding of events in video is common in many human-centric fields such as psychology, sociology, and much of human-computer interaction [3, 4]. This process involves manually searching through videos to find and mark individual events. To avoid errors, only one type of event should be coded at a time [4]. Since the videos are often very long (potentially many hours), and coding the video can take several times longer than the length of the video content itself, the time cost involved can be very high, making it hard for researchers to reanalyze video based on initial findings, or to thoroughly explore the types of events present in their data.

Prior systems such as ANVIL [5] make the coding process easier by providing interfaces for easily annotating audio and video with event tags. However, using ANVIL still requires training, and more importantly, still requires users to watch the entire video multiple times to code more than one event. Our approach uses the parallelism of the crowd to code events in lengthy videos quickly.

Crowdsourcing Video Annotations

Crowdsourcing leverages human computation in the form of open calls to online workers.Crowdsourcing has been used before on tasks that are difficult for automated systems, such as answering visual questions [2], and intelligently controlling interfaces using natural language [6]. The crowd has also been previously used to training activity recognition (AR) systems. For instance, VATIC [9] asked crowd workers to tag objects in a scene. Legion:AR [8] used the crowd to label low-level actions in video by asking workers to watch a live video stream, and then provide an automated system with labels within a short time-frame after they occurred. Unlike our approach, neither of these systems were able to process video any faster than an individual could, and Legion:AR was designed for use not by a human analyst, but by a Hidden Markov Model based system to fill in gaps in its knowledge. With Glance, our goal is for workers to identify more complex events, and to accurately identify the time range over which they occurred.

SYSTEM

Glance has three main components (Figure 1): the analyst interface, the crowd interface, and the merging server.

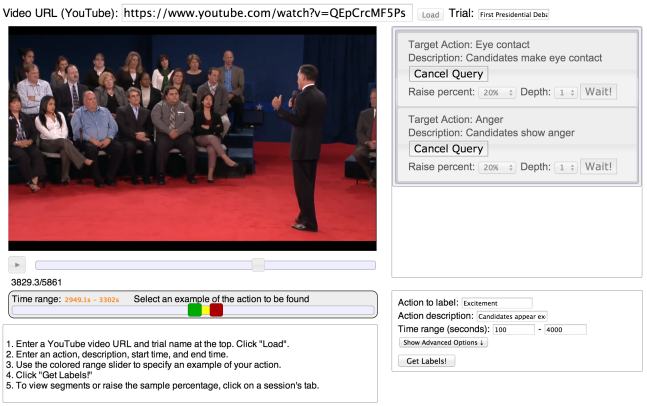

Figure 2. Glance gives analysts a simple to use interface in which they can code multiple events simultaneously, and select how thorough the analysis should be for the current stage of their data exploration.

Analyst Interface

Glance's analyst interface (Figure 2) allows researchers to view their video data, while posting questions about events that might occur within it. Analysts begin by posting their video content YouTube and providing a URL to load it in the viewing area. When an analyst wants to ask a question, they provide a name for the event, a short description, and select a time range to search in (potentially the entire clip). They may also optionally select an example from the video to help demonstrate to workers the event they wish to identify.

Analysts can also select a clip length and sampling rate – the percentage of clips from the whole segment that are issued to workers. They can also set a 'confidence level' to adjust the number of workers who view each clip. Glance then automatically divides the video into clips.

Once a question is asked, a query is sent to workers who mark event occurrences using the worker interface described below. As results come back, the starting points of each occurrence are displayed as markers below the playback bar. Analysts can switch between the results from different queries by selecting the query from a status window displayed to the right of their video viewing window. Queries currently being viewed are highlighted, answered queries are displayed normally, and queries in-progress are partially greyed out (though preview results can still be viewed as they arrive).

Crowd Worker Interface

When crowd workers accept the Glance task, they are shown general instructions and asked to complete an interactive tutorial that verifies they understand how to use the interface. They are then placed in a retainer pool until the task is ready [2, 1]. When a query arrives, workers are routed to a page that shows workers the description of the target event and a video example, if the user provided one. Workers then enter the task and are presented with a video clip and the corresponding set of sliding selectors that they can use to define the start and end time of an event when they see it. As they identify events, new selectors are added to allow them to mark multiple occurrences of the same events in the clip. Workers must watch the entire clip before submitting.

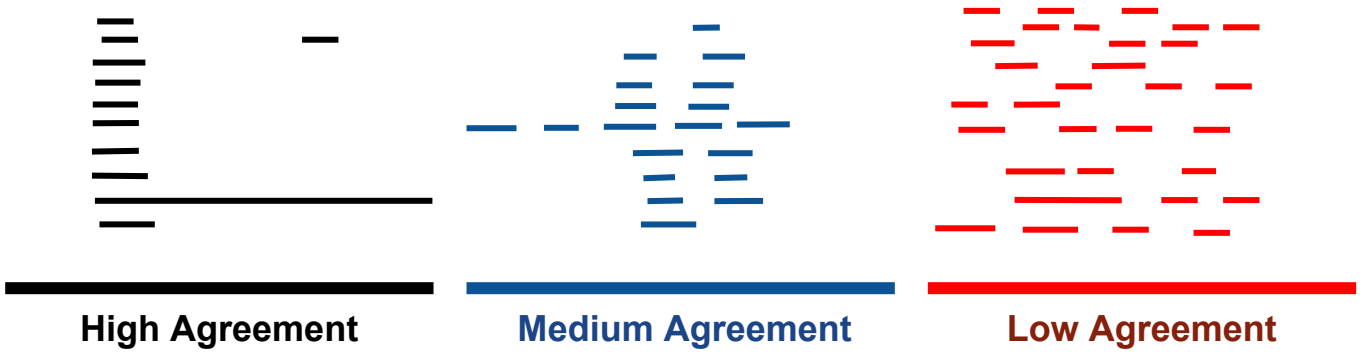

Figure 3. Visualization of the agreement between multiple workers for three 30-second segments. The X-axis is workerID and the Y-axis is segment (color) and time (position). Each bar represents an event marked in the video clip.

Merging Results

Since any single worker completing the task might make an error, multiple workers can redundantly code each clip to increase reliability. In order to combine their results into a single final answer that the user can easily view, we first identify the most likely number of events contained in the clip by taking the mode of the number of events labeled by all workers for that clip. We then filter out *conflicting* events, such as when a worker subsumes two shorter events in a single larger event. Then, we cluster the remaining time ranges using 2-dimensional k-means (where the selected number of events is k). The start and end times of the event ranges within each cluster are then averaged to find the final start and end markers for each event, then displayed to the user.

Feedback to Analysts

Rapid interaction not only changes how the user can interact with data, but also provides the opportunity to change the way the user interacts with the system by giving them more detailed feedback. The patterns of agreement seen in Figure 3 occurred in our tests – using this information, Glance detects when workers might not fully understand the task, and lets the end-user intervene before paying to complete the tasks. Our ultimate goal is to let analysts to have a robust, well-informed two-way conversation with their data - asking questions, and clarifying or updating queries, giving them a more complete understanding of their data than currently possible.

EVALUATION

The goal of our study was to investigate issues of (i) completeness, (ii) speed (latency), and (iii) accuracy of the video coding by Glance. We also investigated how phrasing can affect worker agreement using ambiguous wording. The data set that we used was one hour from the 2012 U.S. Presidential debate between Barack Obama and Mitt Romney. We selected four events to code: two physical events, and two gestalt events. Workers were asked to indicate when the candidates made eye contact, or switched from a seated to standing position (physical), and when they were arguing directly with one another, or their mood changed (gestalt).

A researcher manually coded the start and end times for events in 5 minutes of the video This produced a baseline containing 17 physical, and 16 gestalt events, which we used in order to estimate accuracy. We divided the videos into 30-second clips (10 clips total), and collected responses from 10 Mechanical Turk workers per clip, for each of our 4 events, resulting in 400 total responses.

Precision and Recall

We define *Recall* as the number of events in the baseline that overlapped with an event marked by workers, whereas *precision* is the number of marked events that overlap with some event in the baseline. Figure 4 shows our results for our two physical and two gestalt events respectively, plotted over all possible subsets of workers selected from the 10 we had code each clip and event.

Accuracy of Event Marking

In order to determine the accuracy of event times marked by workers, we measure two things: how far the center of a segment's range is from the center of the baseline range of an event, and the difference in size between the span of the baseline activity and the length of the marked span of the events.

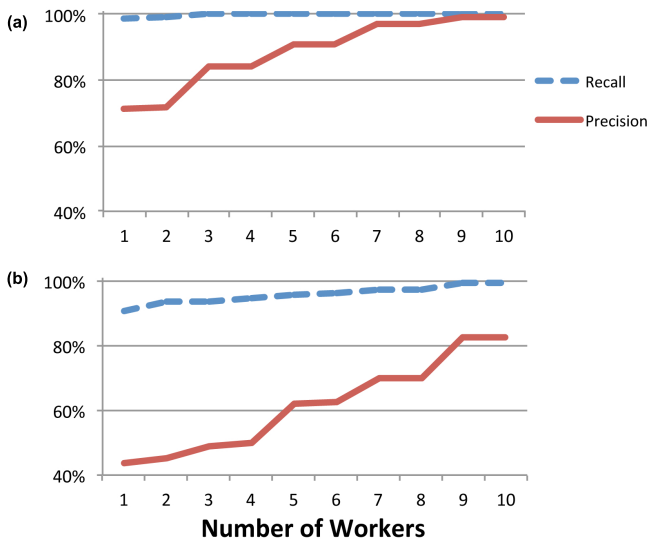

Figure 4. Precision and recall curves. *(a)* Physical events. Precision and recall increase with the number of workers when coding physical events, reaching 99.0% precision after 8 workers. *(b)* Gestalt events. Precision and recall also increased with the number of workers for more gestalt events, but precision only reached 82.3% after 10 workers.

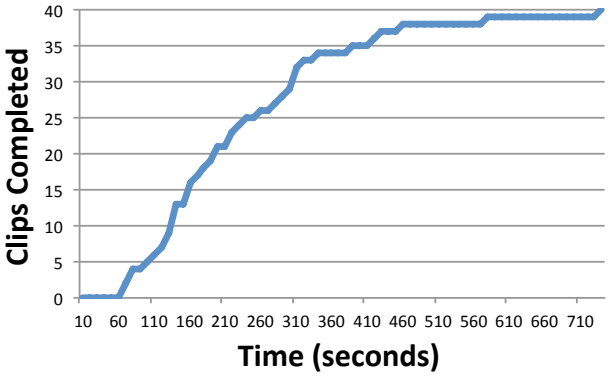

Figure 5. A plot of the segments from an hour-long video being completed by crowd workers in real-time. In two minutes, 20% of the content was labeled. In 5 minutes, 80% was labeled.

This gives us a measure of both the alignment and scale of the segment marked by workers compared the baseline. For the physical measures, the center point of workers' combined event marks averaged just 0.60 seconds from that of the baseline. In terms of length of the events, they were just 0.88 seconds from the baseline range.

Response Latency

An important aspect of Glance's ability to provide analysts with an interactive way to explore their data is response speed. Even an accurate solution is not enough if it still requires weeks to process. We measured the average time that it took workers to view and mark events in our 30-second clips. Workers took an average of of 59.1 seconds ($\sigma = 13.1$) to code the physical events, and 61.0 seconds ($\sigma = 11.7$) to code the gestalt events. There was no significant difference in latency between the two conditions ($p = 0.64$).

In order to confirm that this low per-clip latency translated to low latency in Glance, we ran a live trial on the full hour-long debate video. We included a 5-second example of our events to show the workers during training. The results from this test are shown in Figure 5. After an initial period of delay while workers view the video and mark events, answers rapidly began to arrive. In the first 5 minutes after submitting the query, 48 minutes of the video (80%) had been coded. As the tasks neared completion, there was a decrease in the rate new tasks are completed, due to fewer workers remaining available.

Agreement

To measure the effect that ambiguity had on the response quality, we measured the standard deviation in the number of answers generated by workers when coding the physical events, and the number of answers generated by workers coding gestalt events. We found a standard deviation of 0.12 for the physical events, and 0.94 for the gestalt events (Figure 3). There was a significant difference in the number of disagreements between workers regarding the number of observed events in the two cases ($p < 0.01$).

DISCUSSION

Our results show that Glance is able to use the crowd to label events in video reliably and quickly by using crowds of workers to each code small pieces in parallel. By providing responses within a few minutes, instead of hours or days as pervious approaches have, we make it possible for researchers to refine and update their hypotheses, then ask new questions about their data, all within a single session.

FUTURE WORK

In future versions of Glance, we will expand the interaction from one in which the user provides one-way queries to the crowd to one in which the crowd can collectively provide a response (similar to Chorus [7]). This will let workers confer with users when event descriptions are unclear, before providing a final response. By letting the analysts hold a two-way natural language conversation with their data, we can not only increase accuracy, but provide feedback on specific issues with their queries, such as when something is unclear.

CONCLUSION

This paper demonstrated that Glance can be used to quickly and reliably code videos at a fraction of their playtime. Results indicate that crowd workers are generally reliable, that accuracy improves with redundancy, and that coding is more difficult for gestalt events (as it is for other human coders). A speed test showed that it is possible to code 48 minutes (80%) of an hour-long video in only 5 minutes. This dramatic reduction in speed (from hours/days to minutes) may allow for new kinds of interactive systems that allow analysts to have a real-time conversation with their data.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation awards #IIS-1149709, #IIS-1218209, and #IIS-1208382, and a Microsoft Research Ph.D. Fellowship.

REFERENCES

1. Bernstein, M. S., Brandt, J. R., Miller, R. C., and Karger, D. R. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *UIST 2011*, 33–42.
2. Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., and Yeh, T. Vizwiz: nearly real-time answers to visual questions. In *UIST 2010*, 333–342.
3. Coan, J. A., and Gottman, J. M. *Handbook of Emotion Elicitation and Assessment*. Series in Affective Science. Oxford University Press, USA, 2007, ch. The Specific Affect Coding System.
4. Heyman, R. E., Lorber, M. F., Eddy, J. M., West, T., Reis, E. H. T., and Judd, C. M. *Handbook of Research Methods in Social and Personality Psychology*. ch. Behavioral observation and coding, To Appear.
5. Kipp, M. ANVIL- a generic annotation tool for multimodal dialogue. *Eurospeech 2001*, 1367–1370.
6. Lasecki, W. S., Murray, K., White, S., Miller, R. C., and Bigham, J. P. Real-time crowd control of existing interfaces. In *UIST 2011*, 23–32.
7. Lasecki, W. S., Wesley, R., Nichols, J., Kulkarni, A., Allen, J., and Bigham, J. P. Chorus: A crowd-powered conversational assistant. In *UIST 2013*.
8. Lasecki, W. S., Song, Y. C., Kautz, H., and Bigham, J. P. Real-time crowd labeling for deployable activity recognition. In *CSCW 2013*.
9. Vondrick, C., Patterson, D., and Ramanan, D. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision* (2012), 1–21.