Glance: Using the Crowd to Rapidly Interact with Data

Walter S. Lasecki
ROC HCI, Computer Science
University of Rochester
Rochester, NY 14620 USA
wlasecki@cs.rochester.edu

Mitchell Gordon
ROC HCI, Computer Science
University of Rochester
Rochester, NY 14620 USA
m.gordon@rochester.edu

Steven P. Dow
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
spdow@cmu.edu

Jeffrey P. Bigham
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
jbigham@cmu.edu

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

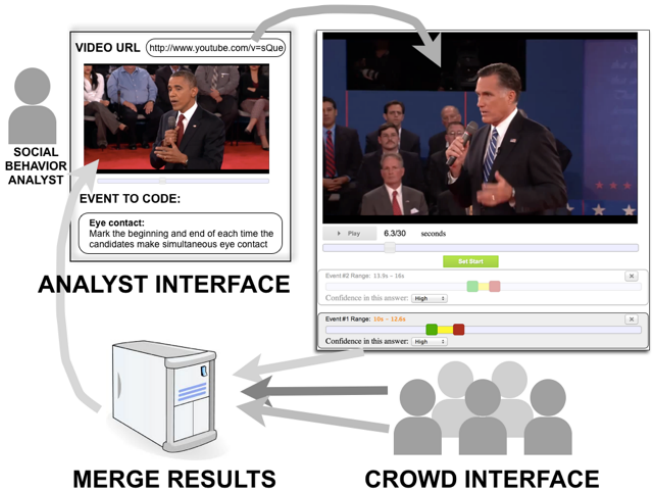
Abstract
Behavioral coding is a common technique in the social sciences and human computer interaction for extracting meaning from video data [3]. Since computer vision cannot yet reliably interpret human actions and emotions, video coding remains a time-consuming manual process done by a small team 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, and 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. 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.

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

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



Analyst Interface
Glance's analyst interface allows researchers to view their video data while posting questions about events that might occur within it. Analysts begin by posting their video content to YouTube and providing a URL to load it in the viewing area. When an analyst wants to ask a question they provide an event name, a short description, and select a time range to search (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, sampling rate, and 'confidence level' (workers per clip).

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 instructions and asked to complete an interactive tutorial that verifies that they understand how to use the interface. They are then placed in a retainer pool [2, 1] until the task is ready. 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 are then presented with a video clip and asked to mark all event start and end times using sliding selectors. Workers must watch the entire clip before submitting.

Merging Results
Multiple workers code each clip to increase reliability through redundancy. 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, which are then displayed to the user.

Future work is exploring new methods for aggregating input that focus on both accurately reconstructing worker responses as well as correcting for worker biases (such as slow response times leading to segment labels that are offset relative to the true occurrence).

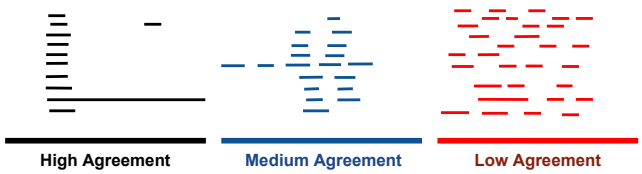


Figure 2: Visualization of the agreement between multiple workers for three 30-second segments. The X-axis is workerID and the Y-axis is segment (color) and time (position). Each bar represents an event marked in the video clip.

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 2 occurred in our tests – using this information, Glance detects when workers might not fully understand the task, and lets the analyst intervene immediately. Our ultimate goal is to let analysts to have a robust, well-informed two-way conversation with their data – giving them a more complete understanding of their data than currently possible.

Evaluation

Prior work in video coding tools has focused on creating easy-to-use systems that allow analysts to code their own video [4]. Crowdsourcing systems, such as Legion [5], have recruited large groups of synchronous workers to complete tasks on a user's behalf. Glance combines work in crowd-powered conversational interaction (e.g., Chorus [6]) with work crowdsourced video labeling [3, 7, 8].

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. We used one hour of video from the 2012 U.S. Presidential debates. 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 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.

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

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

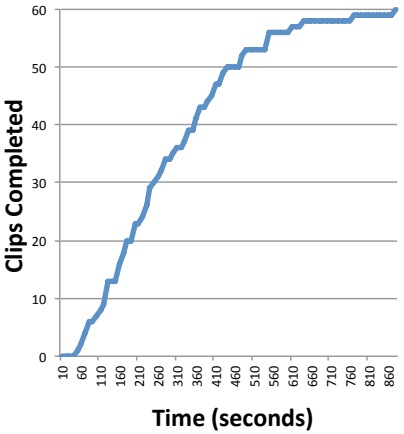


Figure 3: The number of segments from an hour-long video completed by crowd workers over time. In two minutes, 20% of the content was labeled. In 5 minutes, 80% was labeled.

average time that it took workers to view and mark events in our 30-second clips. Workers took an average 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 3. After an initial period of delay while workers viewed the video and marked events, answers began to arrive rapidly. 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 were completed, due to fewer workers being 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 2). There was a significant difference in the number of disagreements between workers regarding the number of observed events in the two cases ($p < 0.01$).

Conclusion

We demonstrate that Glance can quickly and reliably code videos in a fraction of their playtime. Results indicate that crowd workers are generally reliable, accuracy improves with redundancy, and coding is more difficult for gestalt events (as it is for other human coders). We show that it is possible to code 48 minutes of video in just 5 minutes using Glance. This dramatic reduction in speed may allow for new kinds of interactive systems that allow analysts to have a real-time conversation with their data.

Acknowledgments

This work is 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*.

[2] Bigham, J. P., et al. Vizwiz: nearly real-time answers to visual questions. In *UIST 2010*.

[3] Heyman, R. E., et al. *Handbook of Research Methods in Social and Personality Psychology*. ch. Behavioral observation and coding.

[4] Kipp, M. ANVIL- a generic annotation tool for multimodal dialogue. *Eurospeech 2001*, 1367–1370.

[5] Lasecki, W. S., Murray, K., White, S., Miller, R. C., and Bigham, J. P. Real-time crowd control of existing interfaces. In *UIST 2011*.

[6] Lasecki, W. S., Wesley, R., Nichols, J., Kulkarni, A., Allen, J., and Bigham, J. P. Chorus: A crowd-powered conversational assistant. In *UIST 2013*.

[7] Lasecki, W. S., Song, Y. C., Kautz, H., and Bigham, J. P. Real-time crowd labeling for deployable activity recognition. In *CSCW 2013*.

[8] Lasecki, W. S., Weingard, L., Ferguson, G., Bigham, J. P. Finding Dependencies Between Actions Using the Crowd. In *CHI 2014*.

[9] Vondrick, C., Patterson, D., and Ramanan, D. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision* (2012).