Exploring Privacy and Accuracy Trade-Offs in Crowdsourced Behavioral Video Coding

Walter S. Lasecki
Computer Science Department
University of Rochester
wlasecki@cs.rochester.edu

Mitchell Gordon
Computer Science Department
University of Rochester
m.gordon@rochester.edu

Winnie Leung
HCI Institute
Carnegie Mellon University
winniel@andrew.cmu.edu

Ellen Lim
HCI Institute
Carnegie Mellon University
eslim@andrew.cmu.edu

Jeffrey P. Bigham
HCI and LT Institutes
Carnegie Mellon University
jbigham@cs.cmu.edu

Steven P. Dow
HCI Institute
Carnegie Mellon University
spdow@cs.cmu.edu

ABSTRACT

Coding behavioral video is an important method used by researchers to understand social phenomenon. Unfortunately, traditional hand-coding approaches can take days or weeks of time to complete. Recent work has shown that these tasks can be completed quickly by leveraging the parallelism of large online crowds, but using the crowd introduces new concerns about accuracy, reliability, privacy, and cost. To explore these issues, we conducted interviews with 12 researchers who frequently code behavioral video, to investigate common practices and challenges with video coding. We find accuracy and privacy to be the researchers' primary concerns. To explore this more concretely, we used sample videos to investigate whether crowds can accurately recognize instances of commonly coded behaviors, and show that the crowd yields accurate results. Then, we demonstrate a method for obfuscating participant identity with a video blur filter, and find, as expected, that workers' ability to identify participants decreases as blur level increases. The workers' ability to accurately and reliably code behaviors also decreases, but not as steeply as the identity test. This trade-off between coding quality and privacy protection suggests that researchers can use online crowds to code for some key behaviors in video without compromising participant identity. We conclude with a discussion of how researchers can balance privacy and accuracy on their own data using a system we introduce called Incognito.

Author Keywords

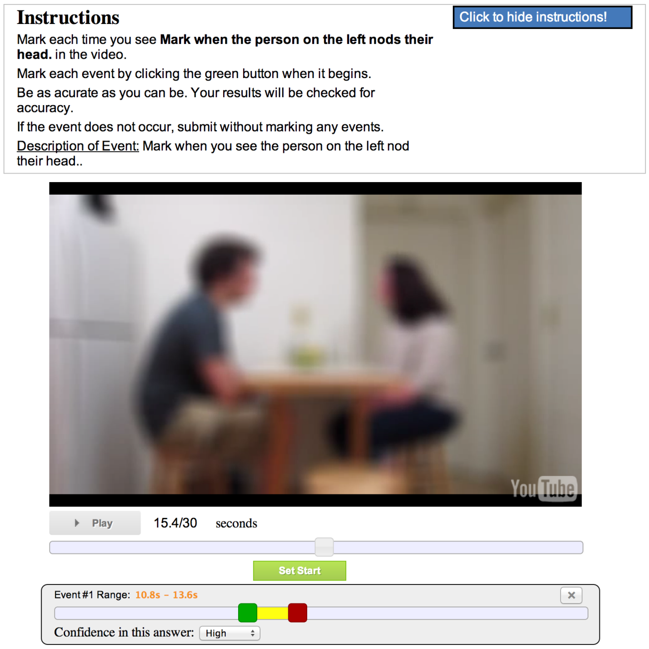
Data analysis; subjective coding; crowdsourcing; video

ACM Classification Keywords

H.5.1 Multimedia Information Systems: Video

INTRODUCTION

Social science and interaction researchers must often interpret video data to gain insight about human behavior in a



Instructions

Mark each time you see **Mark when the person on the left nods their head.** in the video.

Mark each event by clicking the green button when it begins.

Be as acurate as you can be. Your results will be checked for accuracy.

If the event does not occur, submit without marking any events.

Description of Event: Mark when you see the person on the left nod their head..

15.4/30 seconds

Event #1 Range: 10.8s – 13.6s

Confidence in this answer: High

Figure 1: Our studies explore the trade-off between the level of obsfucation and coding accuracy. This figure shows the interface used by workers recruited through Amazon Mechanical Turk to identify behaviors in obfuscated video.

specific situation or interaction. Even with great advances in computer vision, computers cannot automatically identify and interpret the nuances of most human behavior. The most common way to extract meaning from video data is to have one or more human analysts manually *code* it – identify specific events, and mark their time of occurrence. This process is known to be very time-consuming, adding significant overhead to analyzing video data [14].

Recent work has introduced methods for coding behavioral video using online crowds [5, 16, 24, 23]. These systems use online crowdsourcing platforms such as Mechanical Turk to recruit groups of people who can help complete coding tasks with less effort for researchers [23, 24], and in some cases, far more quickly by leveraging the parallelism available on such platforms [16]. However, the shift from using coders who are

CHI 2015, April 18–23 2015, Seoul, Republic of Korea

ACM 978-1-4503-3145-6/15/04 $15.00
http://dx.doi.org/10.1145/2702123.2702605

part of a research team, to untrained, relatively anonymous online workers, presents a new set of potential concerns.

In this paper, we report on interviews conducted with 12 experienced behavioral researchers to understand common practices with video coding, and to uncover their concerns with using crowd-powered systems to code video data. The two primary concerns they had with using the crowd revolved around maintaining the *accuracy* of behavioral judgements and protecting the *privacy of study participants*. To explore this tradeoff, we investigate the accuracy of using online crowds to analyze five commonly coded behaviors. Our results show that crowd workers produce high precision (median 89.5%) and recall (median 78.4%) in comparison to expert-coded behavioral judgements.

We explored how image blurring might be used to address these privacy concerns. We found that for some types of behavioral events, such as touching of the face and body position, privacy can be safely preserved while also maintaining accuracy. Conversely, we found that the accuracy of coding other behaviors, such as eye contact and smiling, dropped when introducing obfuscation techniques. To help researchers explore the tradeoffs of online video coding for their data, we created a software tool called Incognito, which automatically samples obfuscation techniques and attempts to strike a balance between accuracy and privacy for particular data and constraints.

BACKGROUND

Our work explores the intersection of behavioral data analysis for social sciences and crowdsourcing. In this section, we outline the work in these areas that has motivated our use case and made crowd-powered coding of video data possible.

Coding Behaviors in Video Data

In order to rigorously analyze video data, social science researchers in a wide range of fields use behavioral video coding [9, 14]. This process involves training multiple "coders" (usually student research assistants) to identify specific events in a video in a consistent manner [2]. This often takes days or weeks for complex events, and requires determining the specific events of interest in advance, and then generating a corresponding training curriculum. To ensure trained coders are consistent amongst one another, segments of video content are redundantly coded and compared against each other to find the inter-rater agreement [10]. If this is sufficiently high, the coders are trusted to complete other portions of the video independently. To avoid missing event occurrences, practitioners recommend that coders make several passes through the video data and *code one event type at a time*. This increases reliability at the cost of significant additional time required to code a video.

Designing the coding scheme can also be a challenge. Analysts may use a bottom-up approach, in which the coding scheme is constructed from behavioral features observed during careful viewing and re-viewing of the video, or a top-down approach, in which it is derived from theory. Most commonly, some mix of these two methods is used [25]. However, data-driven exploration of video data is difficult using conventional techniques because of the time needed to look for events in large video data sets (consisting of tens or hundreds of hours of video, or more), and the time and monetary cost of recruiting and training a team of coders to perform each stage of the exploratory analysis.

Tools such as ANVIL [15], Datavyu [1], VACA [8], and VCode [12] have been developed to provide interfaces for easily annotating events in video. Despite the availability of such tools, the time required to code behavioral video remains high because it remains a linear process that grows with the total size of the video data set.

Annotating Video Using the Crowd

Crowdsourcing has been used on tasks that rely on human judgment and that are difficult for automated systems. For example, Soylent [3] uses the crowd to edit or shorten writing, VizWiz [4] answers questions about photographs quickly, and Legion [17] follows natural language commands to intelligently control a GUI.

The crowd has also been leveraged in the context of activity recognition systems. For instance, VATIC [24] allows crowd workers to tag where objects appear in a scene. While the crowd provides annotations, it is not designed to respond quickly to the end-user. Similarly, Legion:AR [18] explores crowd labeling of *low-level* actions and even activity structure [20] in video for assistive home monitoring: workers are asked to watch a video stream and provide labels as events happens live; an automated system then learns from these labels for future occurrences. Legion:AR does not process video any faster than an individual can, and is designed for use *not* by a human analyst, but by a Hidden Markov Model-based system.

DiSalvo *et al.* [11] added game elements to an annotation task to get the crowd to mark where an object appears. ”Guess What?” is a game that allows the crowd to help annotate affective behaviors in video. Prior work has shown that using the crowd for affective behavioral coding can collectively be comparable to an expert coder in accuracy [5].

Online crowdsourcing platforms, such as Mechanical Turk, provide a means of easily hiring and distributing small tasks to large sets of crowd workers. Glance [16] uses this on-demand access to a highly parallelizable workforce to dramatically cut the time it takes to get video coded by the crowd to a few seconds or minutes, instead of the hours or days achieved by the quickest previous approaches. This paper extends prior work by investigating how behavioral researchers might leverage the crowd for video coding and what concerns they may have in doing so.

PRACTITIONER INTERVIEWS

While Gottman's techniques provide a basic intro for individual coders, we wanted a better understanding for how behavioral researcher and their teams practically deal with coding video. Also, given the emerging technology for leveraging online crowds for video analysis, we wanted researchers' perspective on the potential benefits and concerns with this approach. To understand if and how crowdsourcing could be

ID	Role	Research area/topics	Behavioral codes
1	Grad student	Rapport in learning settings	Eye gaze, gesture, non-verbal behaviors
2	Faculty	Relationships and stress	Eye gaze, non-verbal behavior
3	Faculty	Group negotiations	Contents of speech (bargaining behavior, reciprocation behavior, etc)
4	Grad student	Rapport & cultural differences	Rapport & cultural differences
5	Faculty	Conflict in groups	Contradictions, aggressiveness, turns in speech
6	Grad student	Classroom culture, turn-taking	Eye contact, gestures, turn-taking
7	Faculty	Couples coping with diabetes	Non-verbal behaviors, emotions
8	Grad student	Communication in person / online	Eye gaze, non-verbal behaviors, rapport judgements
9	Faculty	Learning in classrooms, cultural difference	Deep v. shallow explanations
10	Faculty	Internet, deception	Turn taking, nonverbal gesture, length of hesitation
11	Grad student	Hormonal stress response	Emotions, physical effects
12	Grad student	Rapport between friends versus strangers, self disclosure	Head nods, smiles, eye-gaze, verbal interactions

Table 1: A summary of interview participants, their research areas, and common behavioral codes used in their research.

used as part of a behavioral video coding process, we interviewed twelve researchers with experience in coding video using conventional approaches. All participants have coded at least 100 hours of video data in the past. The researchers comprised 6 faculty members and 6 graduate students from a range of domains across HCI, psychology, and sociology (see Table 1). We recruited participants from an academic university through email and word-of-mouth; no renumeration was offered. Interviews lasted from 30 minutes to an hour. We asked the researchers to describe their current practices around video coding, including how they create their annotation coding schemes, how they record and manage video, and how they analyze video data. We were particularly interested in surfacing common behavioral events and issues that occur while coding video. Finally, we opened a discussion about leveraging online crowdsourcing for video coding and asked participants to reflect on potential benefits and concerns.

All interviews were audio recorded and transcribed. To analyze the data, we went back through the interviews, took down notes, printed out the data, clustered them using open-coding techniques and affinity analysis to find common practices and recurring themes.

Findings

Our findings explore the current state of video coding practice for behavioral researchers and the primary concerns with a crowdsourced approach to this process.

A time consuming process

The current process of video coding is quite time intensive. As P1 claims "it can take ten times the recording time to annotate (the data)." Most researchers reported that the coding process for a particular study can typically take an entire semester or more. One professor commented that "sometimes I can't finish coding in a semester. It might take a year to finish coding one system (P2). Another professor talked about how he collects data during the school year so that he can hire several full-time coders during the summer. Many faculty hire undergraduate research assistants at relatively inexpensive hourly rates, however they report issues with finishing the coding around the students' schedules.

Wide variety of contexts and behaviors

Participants use video coding for a variety of domains and research questions (Table 1). Eight out of twelve participants generally explore research questions about communication and relationships between people, ranging from couples dealing with stress to group negotiations. Two researchers focus on classrooms and group settings; they use video coding because provides a means to "be as unobtrusive as possible" (P9). The final two of our participants look specifically at how people interact with and through technology (e.g., online tutors, Skype calls). Across all the projects, the researchers often coded their video data for non-verbal gestures, facial expressions, eye contact, and turns of speech. To capture these behaviors on video, the researchers typically rely on a single camera positioned so that it can see participants' faces and upper torso.

Developing and refining coding schemes

Behavioral researchers either start with an existing coding scheme or develop one from extant theory. P4 remarked that "the annotation scheme comes out of a mix of two things: the first thing is what does the data look like, what physical data would you code for if you wanted to, and the other is literature, what other people have already." Nearly all participants talked about referencing existing coding manuals from prior literature and tailor them for their own projects. As P3 notes, "we never code for something unless we have a strong motivational hypothesis." Once researchers establish this theoretical grounding, they must develop a coding manual, validate it with their data, and calculate inter-reliability between multiple coders. Sometimes this process never pans out; P3 described trying to code for rapport "we went through (a coding scheme) like six times and eventually decided it was too difficult." For P7, this process "took us probably 6 weeks ... and we started out with a set of codes." The coding manual often goes through many iterations before inter-reliability is achieved, and only then does formal video coding begin.

Preparing people to be video coders

All researchers discussed their processes for training video coders typically undergraduates to use their tailored coding manuals. First video coders read and familiarize themselves with what can often be an extensive coding scheme. One researcher made a "cheat sheet" (P7) so that she could use an abbreviated version of the massive coding manual. P2 discussed how she instructs coders to "watch the video once without doing anything" and then after coding a small portion of the overall data, they "come back and compare with previous coding data and justify ratings." Researchers vary in terms of how they segment and distribute video among a coder team. Several interviewees reported that they typically give the graduate student final authority to meet and resolve disagreements or disparity in coding. After a team of coders achieves inter-coder reliability (typically measured using either Cohen's or Krippendorff's kappa), they can finally begin to individually code segments of video.

Reactions to online video coding with crowds

In general, researchers showed interest in the idea of using online crowds as part of a video coding service, especially if it allowed them to iterate faster on their research questions. P2 discussed how such a service could amenable to "sequential coding" since the video could be broken into chunks. P1 hypothesized that "emotional constructs may be easier for (online workers)." A crowd-based video coding service would need to evaluate efficacy for a range of behavioral events.

Concerns with quality/reliability

In general, the researchers questioned the quality and reliability of data obtained from crowd workers. Many of those interviewed look at non-verbal gestures, which are typically subtle and require either extensive training or a solid grounding in theory. As P6 said "the more interesting gestures are the ones that might have ambiguity." P5 talked about the "subtleties in the way people negotiate" and that coders need to develop "a deeper level of understanding" when coding. P9 noted that she would consider whether online video coders "havent been thinking about these things for a while." Although, P3 noted that "we want people who don't know the theory, because you don't want them to guess the hypotheses and potentially influence their coding." Likewise, P4 mentioned that "annotators did not need to know about the science."

Several researchers brought up the issue of training. Considering the arduous process of training undergrad RAs, P9 wondered how the crowd would be trained:

> If I need ten thousand workers and it takes two hours to train each one, I don't know if you get into issues with the financial costs or the consistency or quality of doing that for each person.

Researchers provided a number of ideas for training. For example, P5 suggested having a tutorial for crowd workers and developing clear, detailed directions for each video. Both P5 and P7 suggested using payment bonuses to reward workers who reach a certain level of accuracy. P6 suggested as a first step to "pick something very easy to code such as eyebrow raises and body movements."

Issues around segmenting data

Some researchers were concerned with how long a clip would have to be for crowdsourcing workers to be able to code reliably. One researcher who does research on conflict and negotiation discussed how "crowdworkers will need to see at least a couple of minutes... the context would be important." (P5) Likewise, a professor noted:

> "An arbitrary time-based unit of chunking wouldnt work because it breaks up the natural units in the thing that you would want to code, they can be speech based or activity based" (P10).

One researcher suggested that such a system could automatically stagger and overlap the clips to avoid this problem.

Concerns with participant privacy and IRB

Additional concerns regarding online video coding revolved around how to protect participants' privacy and whether universities' institutional review boards (IRB) would approve. All participants commented that it would be necessary to obtain specific permission from the participants to upload videos. One participant noted that "The IRB should be okay with this study if the participants know ahead of time that the data would be shared" (P10). However, the IRB needs sufficient evidence that the process is safe and will keep the identity of the participants confidential. P6 suggested that a service would need to "convince the IRB that the process is legitimate and safe (by) first using publicly available data where identity doesn't matter."

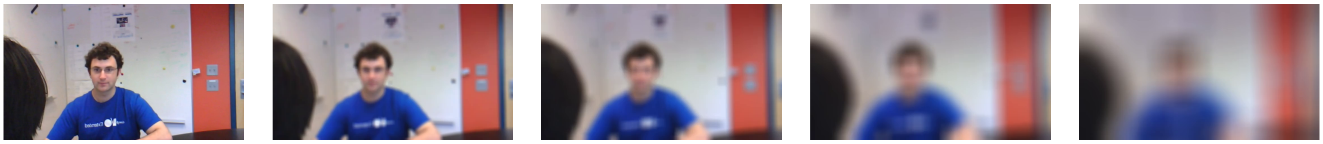
Discussion of Interview Study

The interviews provided greater context around the process researchers go through when coding video data. We learned more about the common strategies and pain points during this process. We also discovered concerns around reliability, video segmentation, and participant privacy that researchers may have with crowdsourcing the video coding process.

We uncovered an interesting potential tradeoff between protecting participant identity and achieving high-quality results. Based on this, we conducted a series of experiments (1) to explore the quality and reliability of using online crowds to code a range of behaviors in video, and (2) to understand if and how obfuscation methods affect the ability to identify behaviors and if these adequately protect participants' identities.

ANSWER QUALITY

The first question we set out to study based on the results of our survey was: "can the crowd accurately answer questions that real researchers want answered?" To test this, we selected the 5 most commonly discussed behaviors mentioned by the researchers we interviewed to have coded by the crowd. For each of the following behaviors, we recorded a video using a volunteer participant who agreed to have their image shared with web workers. Each behavior is listed along with the instruction that we provided to workers:



(a) No Blur (b) Level 6 (c) Level 5 (d) Level 4 (e) Level 3

Figure 2: Examples of the blur levels we used in our experiments. At level 3 almost no personally identifying features can be made out. However, most broad motions can still be made out. As the blur level decreases, more fine-grained features can be identified. Beyond level 6, the effects of blurring are not easily detectable. Level 10 is no blur.

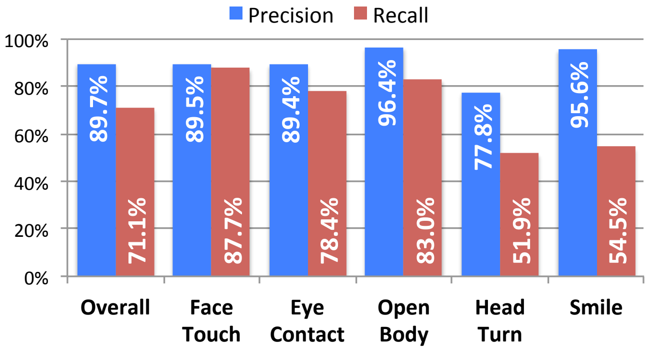


Figure 3: Precision and recall rates for each of the 5 most commonly coded types of events from our interviews with researchers. Note that while the overall scores are high, the `Head Turn` and `Smile` events showed lower recall due to ambiguity in the description given to workers – a common issue for both crowd-powered and traditional coding methods.

- `Eye Contact`: Mark when the person facing the camera makes eye contact with the interviewer.
- `Open Body`: Mark when the person facing the camera has an open body position. This means that they are leaning back in chair/away from evaluator and spreading their limbs apart. [1]
- `Smiling`: Mark when the person facing the camera is smiling.
- `Face Touch`: Mark when the person facing the camera is touching their face.
- `Head Position`: Mark when the person facing the camera turns their head away from the camera/interviewer.

These actions comprise important actions for researchers and span both objective and gestalt, accounting for a wide range of the possible types of actions that researchers may wish to code for with the crowd.

Our interviews with behavioral researchers also revealed that most video is recorded in a lab setting and contains a single person being filmed. The most common camera angle is front-on (often over the shoulder of the evaluator). The side-on angle is also often used, but to limit this study to a reasonable amount of trials, we will focus on the more common front-on angle. As these videos are filmed in a lab setting, they are typically well-lit and high quality. Therefore, we simulated a video recording in a lab setting, and using an actor filmed from a front-on, over the evaluator's shoulder angle.

Experiments

We recruited 89 workers from Mechanical Turk to perform the coding. Figure 3 shows the performance of each of the 5 event types over our video, which was segmented into 6 pieces each with each piece coded by three unique workers. Overall, the crowd averaged a precision of 89.7% (median 89.5%, SD=7.4), and recall of 71.1% (median 78.4%, SD =16.6). Closer inspection shows that not all tasks performed equally well, as would be expected. In this case, the `Head Turn` and `Smile` had 51.9% (SD=43.5) and 54.5% (SD=31.1) recall rates respectively, compared to the other behaviors which averaged 80.1% (SD=4.63) recall between them. This might be due to the fact that the instructions for `Head Turn` and `Smile` are minimal, and while the events seem reasonably intuitive, the ground truth example included even very subtle events that were technically part of the code's definition. The disagreement (low kappa score) between crowd workers at these points likely would mean that in real settings, the system would alert the researcher to the potential ambiguity.

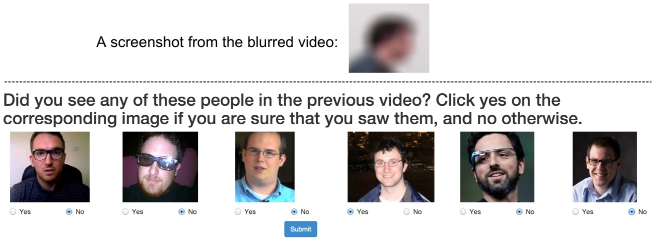
Quality Discussion

One potential reason for this disparity in two of the five cases is that these two behaviors were also the least specifically described of the five. Due to this, workers might have been unclear on a number of the cases that were included in both conditions, leading to the lower recall, while still maintaining relatively high accuracy (77.8% (SD=39.0) for `Head Turn` and 95.6% (SD=8.2) for `Smile`). This reasoning is also supported by prior work that found more vague behavioral descriptions led to less worker agreement and higher error rates.

PROTECTING PRIVACY

Protecting the privacy of study participants was the other primary concern identified by our interview participants. Ensuring this protection is important for running an ethical study with fully informed users, recruiting participants who might be uncomfortable with their likeness being distributed to web workers, and properly informing IRBs of risk.

[1] This is defined in the literature as the face/body leaning back in chair/away from evaluator and spreading their limbs apart.



A screenshot from the blurred video:

Did you see any of these people in the previous video? Click yes on the corresponding image if you are sure that you saw them, and no otherwise.

○ Yes ● No

○ Yes ● No

○ Yes ● No

● Yes ○ No

○ Yes ● No

○ Yes ● No

Figure 4: Our study of participant privacy in crowdsourced behavioral video coding used a lineup tool to test whether or not crowd workers were able to recognize individuals (shown in the bottom line-up) based on the content of the video they were asked to code (an image of the video is shown at the top of this figure, but was not shown to workers).

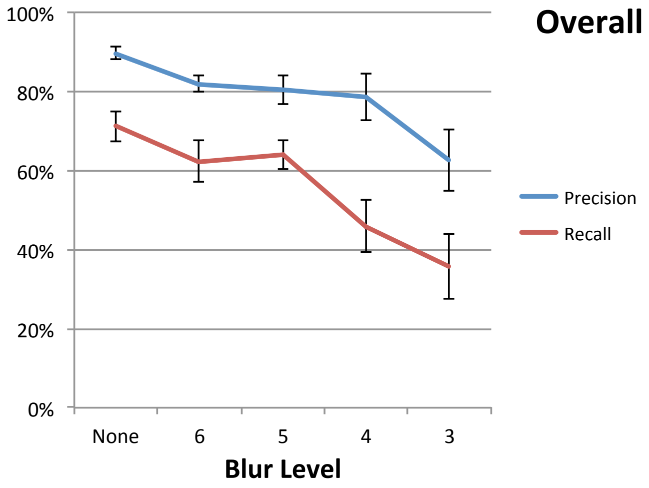


Figure 5: Overall average results from all five of the events we coded for over the unaltered video and 4 levels of added blur. Error bars show the standard deviation. As the severity of the blur filter is increased (from level 6 to level 3), both the precision and recall of the crowd's responses begin to fall.

Related Work

Prior work gives us a means of combating these privacy concerns, though it comes with the risk of reducing the ability of workers to identify certain actions with the same accuracy. These approaches have not directly been focused on preserving privacy in study settings.

Identifying Personal Identity Visually

Gauging people's ability to recognize persons from visual information has been well studied in the past. For example, law enforcement often relies on witnesses recognizing individuals from suspect lineups. The Police Chiefs Association of Santa Clara County published a document containing guidelines as to how to correctly conduct a police-style witness identification lineup [22]. Guidelines include information such as "individuals may not appear exactly as they did on the date of the incident as head and facial hair are subject to change" and "photos/persons will be presented in random order." We used the guidelines in this document when creating our identification lineup tool in order to ensure its accuracy.

Many researchers have studied the accuracy level of these police-style lineups. Henderson et al. conducted multiple studies where participants were first shown a mock robbery CCTV video, then shown photos of the two robbers and seven similar looking people [13]. In one study, they were shown low quality CCTV footage, where the participant's accuracy in correctly identifying the robber was 28%. In a later study, different participants were shown higher quality footage shot by a documentary crew, and their accuracy increased to 76%. In addition, Bruce et. al conducted a study where images of a target were grabbed from a video and shown to participants. Participants accurately identified the target in a video 79% of time when shown a straight on image of the target and 70% of the time when shown an image of the target where the head angle varied by 30 degrees [7]. Both of these studies show that this method is a reasonably accurate way to determine whether someone can be identified visually. Their accuracy rates compare similarly to ours, as we discuss later.

Obscuring Information in Video

Prior work has examined multiple methods for obfuscating people's identity in video while maintaining awareness of actions, such as applying a blur or pixelation filter. Boyle discusses how blurring proved to be more effective at obscuring identity than pixelation [6], and presents 10 different levels of filtering (10 being no blur, 1 being the highest blur level), representing a spectrum of magnitudes for which the effect can be applied. Boyle et al. explore what can protect privacy while letting a large portion of people still identify very course-grain information in an office space, such as how many people or objects are in a room. They do not explore what impact it has on people's ability to recognize events that researchers are interested in when coding video, which often include very subtle behaviors. The office setting used by Boyle et al. is also very robust to errors: there is no critical issue caused by relatively low recognition rates, unlike in the behavioral video coding setting we explore in this work.

In our initial experiments, we chose to replicate Boyle's blur filter and the exact increments of magnitude for each level. Boyle's blur uses a box filter, meaning that a pixel's filtered color is the mean of the neighborhood of pixels surrounding it. Unlike the discrete regions seen with a pixelation filter, the image smoothly changes from one region to the next. Boyle implemented this algorithm such that each level corresponded to a fixed box size that divides into their video's resolution. Our implementation is similar, however it can accept any resolution of video and adjust the box size accordingly.

Prior work has considered how to preserve identity while hiding possibly embarrassing actions in a telecommuting setting [21]. They found that blurring is not an effective privacy tool for home-based video conferencing. This work attempted to hide the actions that were occurring, rather than hide the identities of who is doing them. Our work looks to do the opposite, and so showing that actions can still be identified in a blurry video suggests our approach may be feasible.

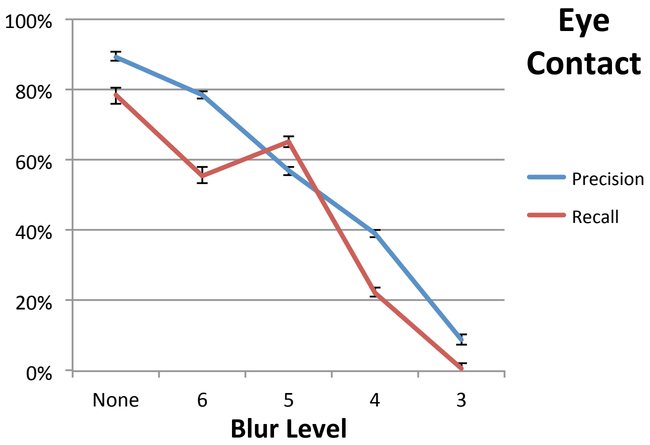


Figure 6: Results from the 'eye contact' event we coded for over the unaltered video and 4 levels of added blur with standard deviation error bars shown. Since identifying eye contact requires recognizing fine-grained movements, both the precision and recall of the crowd's responses fall dramatically as the severity of the blur filter is increased from level 6 to level 3. The 'eye contact' event was the most highly affected by blur of the 5 events we coded.

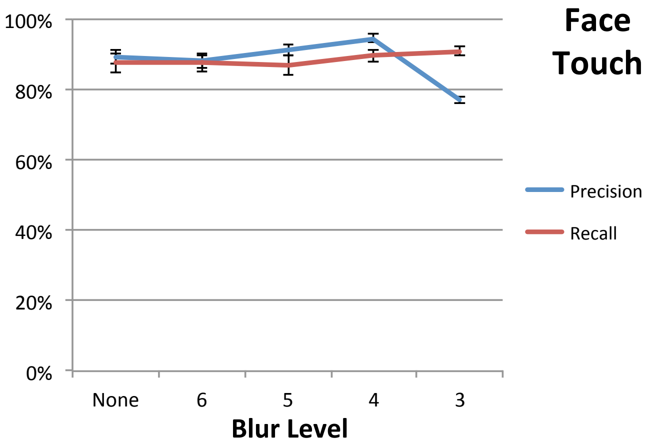


Figure 7: Results from the 'face touch' event we coded for over the unaltered video and 4 levels of added blur with standard error bars shown. Since identifying when someone touches their face requires recognizing relatively drastic movements, both the precision and recall of the crowd's responses are almost unaffected as the severity of the blur filter is increased from level 6 to level 3. The 'face touch' event was the least effected by blur of the 5 events we coded.

Exploring Privacy-Quality Trade-Offs

Since no existing approach can reduce the visual identifiably of participants without also potentially interfering with the identifiably of those participants' behaviors, we want to establish what type and severity of impact privacy filters have on accuracy. We focus on visual privacy in these studies since we found it is most often the focus of researchers' inquiries. To discover how privacy and accuracy can be traded off, we designed our experiment using video of the commonly coded behaviors that we used in our earlier quality experiments.

Experimental Setup

Our study was evaluated using Amazon Mechanical Turk with 553 unique workers. Tasks paid between 19 and 21 US cents. No worker coded a video or completed the facial identification task more than once. This ensured that workers did not become more accurate at identification after having multiple attempts and views of the person to learn from. For example, we want to avoid situations such as letting a worker who coded a video clip with a low amount of blur go on to later code the same clip with a higher amount of blur – the worker would likely be more accurate than they otherwise should have been.

Video Obfuscation Approach

As we discussed earlier in our related work section, Boyle et al. [6] looked at using 10 levels of blur to obfuscate sensitive information in video, and found it to be fairly effective when compared to other techniques such as pixellation. Additionally, blur can serve well as a catch-all for a large variety of actions. Building on this work, we replicated Boyle's levels of blur and chose to code video on levels 3, 4, 5, 6, and 10 (unblurred). We chose these levels because they represented the range of reasonable amounts of blur that would be used with our video. Because our video contained rather large, zoomed in objects, any level of blur higher than 6 would have been indistinguishable from coding an un-blurred video. Likewise, any level of blur lower than 3 would have been too blurry to reveal any interesting results. Figure 2 shows examples of the different blur levels we selected for our tests.

We also ran an exploratory test of a second obfuscation method that exposed only a partial window of the video. Coding for some actions requires only a small, non-identifiable portion of the human body to be visible, e.g., coding for smiling requires that the mouth be visible, but the rest of the video could be heavily obfuscated without affecting accuracy. Masking all but a window could allow sensitive areas to be hidden while retaining accurate coding. We tested this for smiling with an un-obfuscated window around the mouth.

Testing the Visual Identifiably of Participants

To measure how well workers could visually identify the participant used on our videos, we selected 5 images of different people with similar appearances. The image of the participant was chosen to be a clear, direct image, but not one directly from the video workers watched – this prevents higher-level information such as clothing color or type from biasing worker's ability to recognize the participant. Our lineup task interface is shown in Figure 4.

After each worker has finished coding a video segment, they are forwarded to a line-up tool (Figure 4) that asks them to identify the person they saw in the video from a set 6 randomly ordered images. This is similar to common police lineups used to let witnesses fairly identify potential suspects.

For most tests, we did not give workers any warning or indication that they would be completing the lineup task after coding a video. However, since a potentially malicious worker might have the intention of remembering who they saw in the video, we also ran a smaller test to determine whether accuracy increases when workers were warned that they would be completing the lineup tool.

Results

Over all 5 of our coded behaviors, we found that there was a clear linear decreasing trend in both the precision ($R^2 = 0.84$) and recall ($R^2 = 0.90$), as blur level increased (Figure 5). However, each separate behavior we asked the crowd to code had a distinct trade-off pattern.

Figure 6 shows the accuracy curve for our 'eye contact' action, which was the *most* affected by increased blur, decreasing from 89.4% precision (SD=5.61) and 78.4% (SD=10.6) recall to 8.9% (SD=21.7) precision and 0.6% (SD=1.5) recall. Figure 7 shows the accuracy curve for our 'face touch' action, which was the *least* affected by increased blur, going from 89.5% (SD=9.1) precision and 87.7% (SD=12.2) recall to 77.2% (SD=38.3) precision and 91.1% (SD=12.8) recall. Figure 8 shows the accuracy curves for the remaining 3 behaviors we coded.

These accuracy results demonstrate two intuitive but important points: coding accuracy is affected by privacy filtering, and the type of behaviors being coded for impact the size of this effect. Our next goal is to see how workers' ability to identify participants in the video is impacted as the privacy filter is applied at different levels.

The number of workers who are able to identify our participant from a lineup containing theirs and other images quickly falls from 90.9% with no blur applied, to just 31.3% when even a light blur filter is applied, down to near zero recall for higher levels of blur (Figure 9). Note that over many examples, we don't necessarily expect high-blur cases to converge to the rate of random selection (16.7%) because they are able to tell us that they are not sure they recognize the participant in any of the images. Interestingly, we saw almost no fluctuation in the number of false positives (incorrect identifications) marked by workers (2 or 3 in our experiments).

To test whether allowing the worker to know they were going to be asked to identify the participant after their task posed an issue, we ran 2 additional tests which alerted workers that they would be asked to spot the worker from a lineup after their coding task. This is an important effect to measure because malicious workers in future tasks will likely know in advance that they plan to steal the information [19]. We tested this at the two extremes of our prior data to check for any detectable effect: blur level 3, and no blur (level 10). We found the recall rate 85.7% for no blur, and 7.7% for level 3 blur – both of these values are nearly identical to the un-alerted cases. This suggests that preparation is not what is preventing workers from accessing and recalling this information later.

Privacy Discussion

The results from our unfiltered baseline trials closely match the accuracy levels found in [16], where the researchers selected their own set of example videos. This provides additional confirmation that our setup matched the relevant prior work. For effective privacy preservation in a task where we still want to collect data, the most important aspect of this data is that the decay rates between the identity recall rates and behavioral coding quality over blur level is not one-to-one. Instead, personally identifiable information is easier to obscure with a simple filter than the behavioral information being coded.

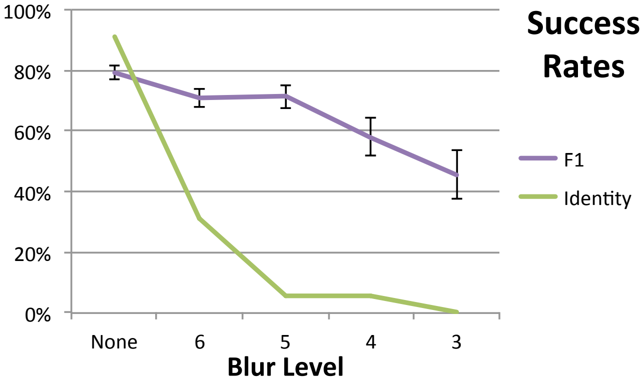


Figure 9: The output of Incognito. The green line (bottom) shows the identity recall rates as the blur level is increased. The purple line (top) shows the average F1 score of the coding results generated by the crowd. Using this, researchers can find the setting-specific optimal trade-off point between privacy protection and coding accuracy for their data.

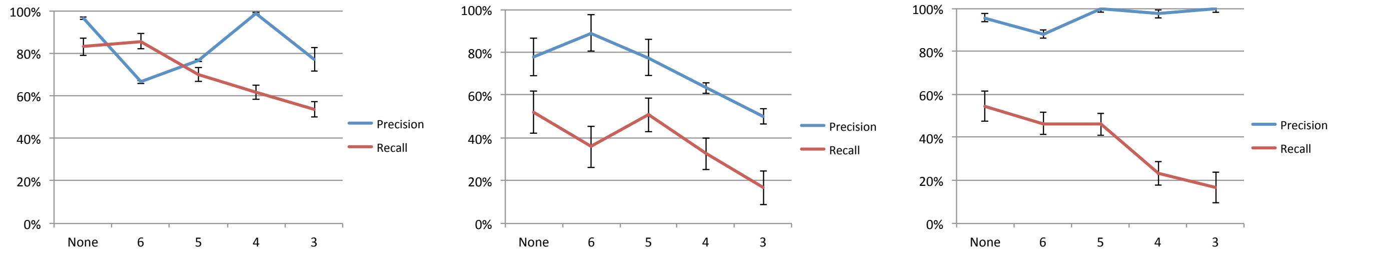
Figure 9 shows both the identity recognition rates, as well as the overall F1 scores[2] of the behavioral coding results. The gap between video coding accuracy and identity recognition accuracy widens as the blur becomes greater. This shows that while it is possible to code for types of actions for which precise vision isn't necessary, identity recognition from a blurry video becomes impossible.

INCOGNITO

Our work aims to demonstrate the viability of crowd-powered video coding in realistic use cases. While we exemplify this with a set of common use-cases, we are interested in making it possible for researchers to guarantee their participant pool that their data will be handled in way that will respect their privacy, even if crowd-based video coding approaches are used, in the specific setting of the research project.

To let researchers explore how privacy may be protected in their specific use cases, we created Incognito. Incognito allows researchers to record example videos of behaviors they would like to code, and then test what level of privacy protection filters are sufficient for their use case and what impact it will have on the accuracy of the crowd-generated coding results. The initial participants used for this calibration can be drawn from a smaller pool of participants willing to share potentially identifiable information, or can be drawn from the researchers themselves.

[2] An F1 score is a commonly used measure that combines both precision and recall into one value. We use it here because it makes it easier to see the trade-off between the two response rate values.



(a) Open Body

(b) Head Turn

(c) Smile

Figure 8: Results from the other three videos in our experiments over increasing blur level, with standard error bars shown. In (c) we see the the results from coding `Smile` differ from the others: while recall falls with high blur levels, precision remains high. This suggests that the types of instances of smiling that are being obscured by blur are limited to a subtle subset, where more obvious examples, *e.g.*, when a wide smile shown teeth, are still discernible by crowd workers.

Usage

To use Incognito, a researcher first provides a set of example videos that show the events and behaviors they wish to code. Incognito then applies a video filter, and then uses our worker interface (Figure 1) to test the crowd's ability to find each event using the filter. Crowd workers can mark one or more segment of the video as containing a specified action, and submit their task when they have finished.

If the filter used can be applied in multiple ways or at multiple intensity levels (such as level of blurring, or mask padding), Incognito can measure the effect of each level and provide researchers with a trade-off curve in terms of precision and recall rate (or F1 score, as shown in Figure 9) over filter level. To ensure that the results are not biased by workers seeing the same content form a prior example (especially important since Incognito will typically operate on a small set of preliminary data), the system recruits a distinct group of workers for each trial it runs.

Reducing Identity Recollection

After each worker has finished coding a video segment, they are then forwarded to the line-up tool (Figure 4), just as in our study. The set of images use in the lineup must be provided by the researcher and should contain roughly 3-6 non-participant images, and then 1 image of the participant that is not directly drawn from the video that workers will see.

Based on the overall identification rate, Incognito can determine if privacy will be sufficiently preserved (based on a specified privacy requirement) at a given level, or that more levels need to be run. Alternatively, the determination will be left to researchers, in which case the final information on event versus identification recall rate and accuracy are provided to researchers to inform their decision (Figure 9). This information can help researchers determine how to discuss privacy issues with their review board and participants.

Analyzing Results

To read the output of Incognito, researchers can find the maximum level of potential privacy risk they are willing to accept *e.g. to satisfy the requirement "no more than 5% of workers should be able to identify any of the participants from video."* on the X-axis, then check the accuracy results on the Y-axis to see if the crowd can perform well enough to produce usable output for the type of video data the researcher intends to code. For example, in Figure 9, blur level 4 would prevent all but around 5% of workers from being able to identify participants, but it will not work if the researcher had to catch at least 2/3rds of all instances of the behavior in question – in that case they could only use level 5 or higher. However, from the results, they would also know that the change from blur level 4 to blur level 5 only results in an additional 0.22% of workers being able to identify participants in the study. In exchange, overall coding accuracy (in terms of F1 score) jumps 13.2%, from 58.0% to 71.2%. Thus, in this case, the minor decrease in privacy level is probably worth the increase in speed and accuracy.

FUTURE WORK

In the future, our results motivate a wide range of work related to how crowdsourcing approaches can be used in real behavioral data analysis settings.

Other Approaches to Filtering Information

As we discussed in the Related Work section, there are many ways to obfuscate video content. While we studied blurring because it is general purpose, easy to implement, and easy to replicate in future studies, we are also interested in the effects other filters have.

For example, we also collected data using a masking filter that completely hid all but a given section of the video a researcher believes they need for their event. We masked all but the participant's mouth to code for `Smile` events in the data, achieving 73.2% precision and 50.3% recall. While these results are lower than the ones we saw when using the blur filter, it speaks to the importance of visual context for workers. Future work will investigate factors such as this, which remain to be fully explored in the context of crowd-powered behavioral video coding.

Real-World Usage Studies

Our results suggest that participant privacy can be preserved well enough to begin considering crowd-powered approaches for use in real settings. In future work, we aim to help facilitate and study the use of crowd-powered systems in the video coding process, and compare the resulting practices with those we explored in our interviews in this work. To this end, we will be releasing our Incognito tool to help researchers find the balance between privacy and accuracy issues that works for their specific needs.

CONCLUSION

In this paper, we explored the privacy and accuracy trade-offs in crowdsourced behavioral coding of video. Through interviews with 12 researchers experienced in coding behaviors, we identified many common themes and types of behaviors that are coded for across different settings and tasks. We also found that the primary concerns researchers had with crowd-powered approaches to video coding were the response quality and protecting the privacy of participants. To investigate these issues, we ran experiments with 636 crowd workers recruited from Mechanical Turk. Our first experiment demonstrated that the crowd is able to answer the 5 most common types of behaviors that are coded with nearly 90% accuracy. We then discussed methods for protecting participant privacy in video, and ran a set of experiments to show the trade-off that occurs when video quality is degraded to help prevent workers from being able to identify participants. Our results show that for many types of behaviors there are opportunities to protect participants identity almost entirely, while still being able to get highly accurate answers from the crowd. Finally, we introduced Incognito to let researchers explore trade-offs between accuracy and privacy in their own data.

ACKNOWLEDGEMENTS

This work was supported by National Science Foundation awards #IIS-1217929, #IIS-1149709, #IIS-1208382, #IIS-1217096, Google, an Alfred P. Sloan Foundation Fellowship, and a Microsoft Research Ph.D. Fellowship

REFERENCES

1. Datavyu. http://datavyu.org/.
2. Bakeman, R., and Gottman, J. M. *Observing interaction: An introduction to sequential analysis*. Cambridge University Press, 1986.
3. Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. Soylent: a word processor with a crowd inside. In *UIST* (2010), 313–322.
4. 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.
5. Borsboom, B. Guess who?: A game to crowdsource the labeling of affective facial expressions is comparable to expert ratings, 2012.
6. Boyle, M., Edwards, C., and Greenberg, S. The effects of filtered video on awareness and privacy. In *CSCW* (2000), 1–10.
7. Bruce, V. Fleeting images of shade: Identifying people caught on camera. *The Psychologist* (1998), 331–338.
8. Burr, B. Vaca: A tool for qualitative video analysis. In *SIGCHI EA* (2006), 622–627.
9. Coan, J. A., and Gottman, J. M. *Handbook of Emotion Elicitation and Assessment*. Series in Affective Science. Oxford University Press, 2007, ch. The Specific Affect Coding System (SPAFF).
10. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas 20*, 1 (1960), 37.
11. Di Salvo, R., Giordano, D., and Kavasidis, I. A crowdsourcing approach to support video annotation. In *VIGTA* (2013), 8:1–8:6.
12. Hagedorn, J., Hailpern, J., and Karahalios, K. G. Vcode and vdata: Illustrating a new framework for supporting the video annotation workflow. In *AVI* (2008), 317–321.
13. Henderson, Z., Bruce, V., and Burton, A. M. Matching the faces of robbers captured in video. *Applied Cognitive Psychology* (2001), 445–464.
14. 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*. Cambridge University Press, 2014, ch. Behavioral observation and coding.
15. Kipp, M. ANVIL- a generic annotation tool for multimodal dialogue. *Eurospeech* (2001), 1367–1370.
16. Lasecki, W. S., Gordon, M., Koutra, D., Jung, M. F., Dow, S. P., and Bigham, J. P. Glance: Rapidly coding behavioral video with the crowd. In *UIST* (2014).
17. Lasecki, W. S., Murray, K. I., White, S., Miller, R. C., and Bigham, J. P. Real-time crowd control of existing interfaces. In *UIST* (2011), 23–32.
18. Lasecki, W. S., Song, Y. C., Kautz, H., and Bigham, J. P. Real-time crowd labeling for deployable activity recognition. In *CSCW* (2013).
19. Lasecki, W. S., Teevan, J., and Kamar, E. Information extraction and manipulation threats in crowd-powered systems. In *CSCW* (2014).
20. Lasecki, W. S., Weingard, L., Ferguson, G., and Bigham, J. P. Finding dependencies between actions using the crowd. In *SIGCHI* (2014), 3095–3098.
21. Neustaedter, C., and Greenberg, S. Balancing privacy and awareness in home media spaces. In *UBICOMP* (2003).
22. Police Chiefs Association of Santa Clara County. Line-up protocol for law enforcement.
23. Riek, L. D., O'Connor, M. F., and Robinson, P. Guess what? a game for affective annotation of video using crowd sourcing. In *ACII* (2011), 277–285.
24. Vondrick, C., Patterson, D., and Ramanan, D. Efficiently scaling up crowdsourced video annotation. *Int J of Comput Vision* (2012), 1–21.
25. Weingart, L. R., Olekalns, M., and Smith, P. L. Quantitative coding of negotiation behavior. *Int Negotiation 3*, 9 (2005), 441–456.