Crowdsourcing and CCTV: the Effect of Interface, Financial Bonus and Video Type

Paul Dunphy[1], James Nicholson[2], Vasilis Vlachokyriakos[1], Pam Briggs[2], and Patrick Olivier[1]

[1] Culture Lab, School of Computing Science, Newcastle University
[2] PACT Lab, Northumbria University

Abstract. Crowdsourcing has been widely leveraged for the tagging of video material; recently this has included the monitoring of surveillance video footage. However, the relative advantages and disadvantages of crowdsourcing watchers of surveillance video are not well articulated, nor has there been any significant work on the efficacy of different tools for retrospective surveillance. In this paper we explore factors that might affect crowd performance on video surveillance monitoring tasks. We firstly established a baseline for crowd performance in a 'live' surveillance study before comparing two different interfaces that allowed retrospective discovery of the same events (Scrub Player and Panopticon). Using MTurk, we asked 474 people to monitor two types of CCTV footage containing different levels of extraneous activity for 'events' using these two interfaces. We also manipulated bonus payments. We found that the crowd was most accurate when watching the video that contained the least extraneous activity, that there was no effect of financial incentive on the reporting of true events, but that there was some indication that higher bonus payments generated more false alarms. We also found an effect of interface: those using the Panopticon interface processed more video footage and generated more alerts, but were less accurate overall. We discuss the implications of different interface designs, video types, and incentive schemes when crowdsourcing watchers of surveillance video.

Keywords: Crowdsourcing; CCTV; Surveillance.

1 Introduction

Surveillance is a subject that fuels debate, particularly in technologically advanced societies, where citizen behaviours can be monitored by a highly diverse network of systems and devices. Closed Circuit Television (CCTV) is often considered the archetypal surveillance technology, used for both the deterrence of crime and as a means of providing evidence in criminal prosecution. This technology is used worldwide, however, countries such as the United Kingdom appear to be peerless in commitment to CCTV; some estimates suggest that 4 million cameras are deployed there [21]. The public perception of CCTV as an omniscient force in society is thought to mask an unwieldy organisational machine that is highly dependent upon a technology infrastructure that is costly to maintain, and upon security staff paid to watch the video footage: generally low paid, under-appreciated, overloaded with other tasks, and fighting extreme levels of boredom [25].

A number of initiatives have been proposed to reduce the reliance upon human labour for event detection in live or archived surveillance video, with many of these involving sophisticated video analytics or machine learning techniques [19]. In parallel, web platforms have emerged that incorporate crowdsourcing, or human computation [29], in the analysis of live footage streamed online. Those recruited are asked to watch surveillance footage and report any notable events to a relevant authority via the user interface. The principal assumption underlying this approach is that a large, spontaneous, and distributed workforce can analyse footage as effectively as a dedicated team of professionals. A number of high profile web platforms work in this way to achieve different goals, for instance, monitoring for evidence of illegal immigration via a video feed from the US/Mexico border [27] as well as routine anti-theft surveillance in grocery stores in the UK [13]. Yet, little is known about the deployment of crowdsourcing platforms in this space, nor about the design and reward factors that might influence the performance of a remote and distributed workforce.

In this paper we attempt to understand more about the event detection capability of the crowd across both live and archive video surveillance tasks. There has been relatively little crowdsourcing work on the latter, despite the fact that the monitoring of archived surveillance footage is a huge problem in crime detection [19] and its absence in the research literature is particularly surprising given the established use of crowdsourcing in the annotation of archived video [31]. We seek to redress this problem in a study in which we first establish baseline performance in a live surveillance task and secondly, compare two platforms for video archive surveillance under a number of different conditions. Our specific contributions are:

- An investigation of the performance of a spontaneously recruited online community asked to monitor surveillance video, conducted under a set of manipulations that allow comparison between (i) different video types (live vs archived; simple vs complex); (ii) different interfaces for video viewing and navigation and (iii) different incentive schemes (variable rates of pay).
- The application of a video surrogate interface in a more strenuous surveillance video monitoring context than had previously been attempted (Panopticon [14]).

2 Related Work

2.1 Understanding CCTV

In the UK, deployment of CCTV in public spaces is thought to have been first adopted as a political tool to provide a visible response to worrying levels of crime. Its actual efficacy as a crime prevention tool has been disputed [7] as there are a number of confounding factors that make it impossible to assume a simple relationship between CCTV deployment and the overall crime rate [2]. Nonetheless, as Tullio et al. [30] note, CCTV typically forms part of a complex evidence system in which the over-riding legal need is to be able to make more effective use of the footage that is captured. This is problematic given the volume of data generated by cameras. Indeed, it has been suggested that our ability to gather video surveillance footage has not been matched by the ability to process all the data that is generated [22]. Surette [28] points out that a typical 20 camera setup that records 24 hours per day will generate 480 hours of video, and 43 million images per day. The appropriate and timely analysis of this material is an onerous task for control room staff.

A number of ethnographic studies have aimed to understand the context of work for CCTV operators; such studies show that the control room is a complex workplace that can, in itself, affect the intended security gains. Smith [26] carried out an ethnographic study of security staff at an educational establishment in the UK and found that the security personnel felt bored and undervalued. They were also seen to adopt practices to combat boredom – practices that conflicted with their need to judiciously monitor visual displays for suspicious behavior (e.g. focusing cameras on their own vehicles, playing hide and seek) [25]. In a further ethnographic study that explored the relationship between watchers and watched in the control room, Smith noted that watchers can develop a perceived rapport with individuals they have never met, but repeatedly see on-screen and develop empathy with them over the longer term [26]. Given some of the problems outlined here with the use of computational video analytics within the legal system, it not surprising that there is a move to consider crowdsourcing as one cost-effective solution to processing the large volumes of data gathered. In the sections below, we give some background to the use of crowd-based classification of video material, including considerations of the design factors that maximise crowd performance, before turning to some examples of the use of crowdsourcing in video surveillance.

2.2 Improving the Performance of the Crowd

There has recently been a move towards a greater understanding of crowdsourcing, with the recognition that both worker incentives [5, 23, 20, 16] and task interfaces [8] can have a significant impact upon participation rates and performance from workers sourced online.

Bespoke interfaces can optimise the performance of the crowd across a range of task types, such as video annotation [30, 31], short video filtering [3], speech captioning [17], and even word processing [2]. The interfaces most pertinent to the task of crowd surveillance of CCTV footage are those that have been developed for video annotation. These interfaces typically use a simple timeline interface with additional features to support the annotation process (e.g. predetermined tag displays, ability to mark objects on the video, etc.). However, to the best of our knowledge, no studies have explicitly addressed the design and/or performance of different video navigation interfaces in a task that asks crowd workers to search for events within video material. That said, we can learn from Lasecki et al. [17] who demonstrates that allowing the crowd workers to control the media (in this case the speed of the audio) significantly improved their task efficiency.

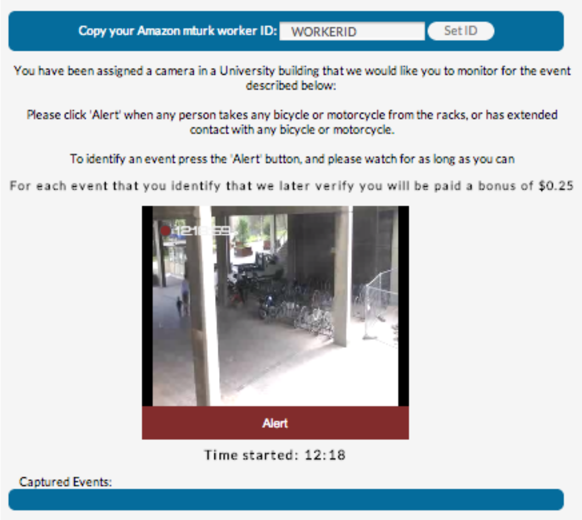
The literature on incentives for crowd workers is more contradictory. For example, Kazai [16] reported that a straightforward financial bonus improved the quality of participants' work – similar to [11]'s findings. In contrast, [20] found that financial incentive increased participation, but did not improve the quality of the work done. Note too, that new work on intrinsic motivation (e.g. framing a task as a social benefit or an altruistic act) suggests that this may be more effective than pay in actually improving output quality [22].

2.3 Crowdsourcing and Video Surveillance

The Texas Border Watch [4] program invited watchers from all over the world to monitor live surveillance feeds from one particular stretch of the border between the United States and Mexico. There were known problems with drug trafficking and illegal immigration on that border and so local residents were invited to place cameras on their own property to contribute a video feed to a central website. Alarms raised by the crowd on the user interface were then sent to an appropriate border Sheriff. During deployment, 26 arrests were made and around 7,400 pounds of narcotics were seized in two years, which worked out at a cost of $153,800 per arrest based on the initial financial investment [10]. Internet Eyes [13] was a private company in the UK that aggregated CCTV from grocery stores and distributed the feeds on their web platform. Each alarm raised by users on the user interface would be sent directly to the management of the grocery store. Watchers were offered the promise of financial rewards if they were able to spot incidences of shoplifting. To preserve privacy, the site claimed to ensure that users would not be able to watch CCTV video from their local area, nor discover where the feed was based. This site appears quite unique in its provision of a particularly voyeuristic experience; the service has an active presence on social networking sites, where watchers discussed recent events seen on camera.

The Shoreditch Digital Bridge [18] was a government funded project in 2006 based in London, UK. This transmitted feeds from 11 local CCTV cameras directly to the televisions of local residents. The feed was accompanied by a ”rogues gallery” which comprised a picture slideshow of individuals in the area with anti-social behaviour orders attached. Residents were not provided with a means to interact with the videos, and any alarms that needed to be raised had to be done so by telephone to the police. This configuration also has a particular voyeuristic appeal as residents were given license to monitor their local communities. Indeed, local reports of graffiti and vandalism were increased by 600% and 200% respectively. It is not clear whether the local constabulary had the capacity or capability to respond to these additional reports of criminal behaviour. The project attracted a number of privacy and civil liberties concerns.

Our own study was motivated by a desire to understand more about the factors affecting crowd participation, engagement and performance within a surveillance context. We carried out two experiments: the first configuration comprised a fairly standard deployment where watchers are presented with a continuous feed of seemingly live video; the second explored the efficacy of two different interfaces, both designed to improve performance in video event-detection. In each study we investigated two further factors, manipulating the complexity of the video watched (in terms of number of distractor events taking place) and the level of financial incentive offered to the workers. On the basis of related work on video search, we believed: (i) that participants would find it difficult to sustain attention and interest in the process of monitoring live video footage; (ii) that event detection was likely to be easier and more accurate in the simple video task; (iii) that reward may lead to workers spending more time on task but would



Copy your Amazon mturk worker ID: WORKERID

You have been assigned a camera in a University building that we would like you to monitor for the event described below:

Please click 'Alert' when any person takes any bicycle or motorcycle from the racks, or has extended contact with any bicycle or motorcycle.

To identify an event press the 'Alert' button, and please watch for as long as you can

For each event that you identify that we later verify you will be paid a bonus of $0.25

Time started: 12:18

Captured Events:

Fig. 1. The user interface for Experiment 1, where the surveillance video is presented as being a live stream. Participants were asked to click ’Alert’ once an on-screen occurrence that met the specified criteria was observed. Video frames that the user selected as being significant were displayed along the bottom of the screen.

not necessarily improve performance and (iv) that there would be an effect of interface, with the new Panopticon system being more effective in facilitating the process of finding events of interest, leading to superior surveillance performance [14].

3 Experiment 1 - Live Video Surveillance

The first user study addressed real-time, seemingly live, surveillance. This configuration is comparable to how the majority of the exemplar platforms have been instantiated. We know that watching video footage in real time can be tedious, and so our research aim was to understand how variations in financial reward and video type might affect performance. We used Amazon MTurk as the platform, as it provided us with a suitably wide user base.

4 Method

Our first study was a 2 (video type) x 3 (incentive level) independent design and the dependent measures were the amount of video watched, user precision and recall, and a complementary measure of false positive interactions.

4.1 Task Materials

We constructed two videos to resemble CCTV footage for the basis of our studies (see Figure 2). These two videos were chosen to represent two quite different surveillance

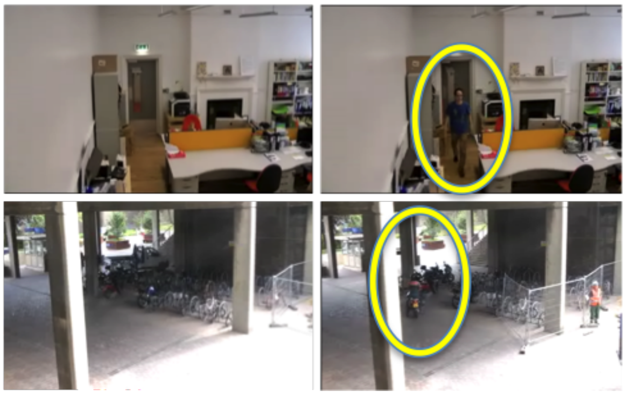


Fig. 2. Extracts from the two surveillance videos that we constructed. Top: a lab context where the images were mostly static providing a particularly tedious viewing experience; bottom: the university bikesheds, a busy area that is already heavily covered by CCTV. Screens on the right side illustrate example events where watchers should raise an alarm (yellow circles are illustrative).

tasks that appear to conceptually represent those seen in our earlier review of existing platforms of this nature, according to the level of activity they contain:

- **Lab video**: Participants were asked to monitor the entrance to a laboratory and asked to raise an alarm when *any* person walked through the door. This relatively static footage relied purely upon a reactionary response from the user. No extraneous events occur. We refer to this as a *simple* video.
- **Bikes video**: Participants were asked to monitor an outdoor area of high public traffic on the university campus that was already heavily monitored by CCTV cameras. On the footage, people can be seen to regularly add and remove bikes from the bike shed or simply walk by. Participants were asked to raise an alarm whenever a bike was legitimately taken or returned to the bike shed. This task therefore required a more nuanced interpretation of the behaviour of the people on-screen. We refer to this as a *complex* video.

The video type likely plays the most important role to influence a number of performance parameters for watchers, and we hoped that the differences between our two videos could be sufficient to help us to gauge the significance and focus of this influence. There is currently no empirical data to guide our intuition in this regard for a surveillance context.

While we presented the scenario that the viewed video was a live feed, both videos were 2 hours long and contained the same number of true positive actionable events (20) with the distribution approximately balanced through editing. The videos contained no sound and their natural size was 352x288 pixels. We assigned each ‘event’ in the video a time period within which a click would be accepted as a useful alarm for purposes of assessing the accuracy of watchers. Our study design passed ethical review and there were no special measures that we were required to implement.

4.2 Procedure

We firstly posted a HIT on the MTurk website advertising a video monitoring task. There we outlined the basics of the study and the criteria for payment. All participants who watched 15 minutes of video would qualify for a small base payment (to provide criteria for MTurk work completion). Once the participant decided to accept the task they were redirected to a website on our own web server, at which point we randomly assigned one of the three incentive conditions: no incentive – participants could only earn a small base payment ($0.40) no matter how hard they worked, low incentive – participants could earn $0.05 for every event that they identified that we later verified, or high incentive – where participants could earn $0.25 for every event they identified that we later verified. The participants were not made aware of other bonus or reward schemes other than the one to which they had been assigned, and were not told how many events were present in the video.

Before viewing the video, participants were also provided with a simple visual example of an event that they were searching for (that resembled Figure 2). We invited them to watch as much of the 2 hour video as they wished. We developed a simple web-based interface resembling the Texas Border Watch website that provided users with instructions, the video feed, and a button underneath that read “Alert” (see Figure 2). We asked participants to ”watch the video and report any events they felt were relevant by pressing the alert button situated below the player window”. The definition of an “event” was always present on-screen and differed according to the video type, along with information of any bonus entitlement. Simple steps were taken to present the impression this was a ‘live feed’: firstly, no playback controls were present on the interface, which also masked any video length detail on the interface, secondly, a clock (to give an impression of local time) was overlaid onto the video feed, and thirdly, cookies remembered playback positions for the case that participants navigated away from the site during the task (which was logged).

4.3 Participants

In total, 114 participants completed this phase of the study and were included in the main analysis. The composition of the recruited participants was 65% male and 35% female. There were a number of participants who started the study but who did not watch the required 15 minutes of video to receive the baseline payment and were not included in the main analysis (unless otherwise stated). There were 27 participants who did not watch the required 15 minutes for the simple video task, and 30 for the complex video task.

4.4 Results

The analyses are presented in the way of independent 2x3 ANOVAs for the main comparisons. We found that the data was not normally distributed, but there was enough variance in the data to mirror the results from non-parametric tests while also capturing any possible interaction effects. We present the results for time engagement, and user accuracy. In the former, our significance testing was focused upon the amount of video analysed by participants, and in the latter this was focused upon accuracy as described by an F_1 score. Other measures are also given to provide descriptive statistics.

4.5 Time Engaged on Task

The prime engagement measure was the time spent by participants watching their assigned video. This is the only analysis that includes those participants who completed less than the required minimum of 15 minutes on task. A 2 (video type: simple, complex) x 3 (incentive: none, low, high) ANOVA revealed no main effect of video type, ($F(1, 108) = 1.108, p = 0.295$), and no main effect of incentive $F(1, 108) = 1.978, p = 0.143$). We observed no interaction between video and bonus ($F(2, 108) = 1.365, p = 0.260$). This suggests that the time spent on the task was neither influenced by complexity of the video itself nor the bonus structure.

Table 1. Descriptive statistics of the accuracy and time engagement recorded from participants in Experiment 1. The lab video in the 'live' study is represented in the top table, and the bikes video in the bottom table.

Bonus	n	Mins of Video	Precision	Recall	F_1
		(μ)	(μ)	(μ)	(μ)
$0	19	16	0.92	0.63	0.73
$0.05	19	24	0.85	0.84	0.8
$0.25	19	25	0.7	0.67	0.71
Total	57	22	0.85	0.71	0.75

Bonus	n	Mins of Video	Precision	Recall	F_1
		(μ)	(μ)	(μ)	(μ)
$0	18	16	0.76	0.7	0.73
$0.05	19	16	0.56	0.62	0.61
$0.25	20	26	0.58	0.52	0.54
Total	57	19	0.63	0.62	0.63

4.6 Accuracy

Table 1 presents a summary of participant accuracy using metrics common to information retrieval: precision and recall. Precision reflects the number of events correctly identified as a proportion of the total number of events identified (both true hits and false alarms), whereas recall reflects the number of true events identified as a proportion of the total number of true events available on the video. The harmonic mean of these two measures produces an F_1 score [3] - where an F_1 score of 1 indicates best possible performance and a score of 0 indicates worst possible performance. Using these

[3] The F_1 score can be considered as a weighted average of precision and recall. $F_1 = 2(\frac{precision*recall}{precision+recall})$

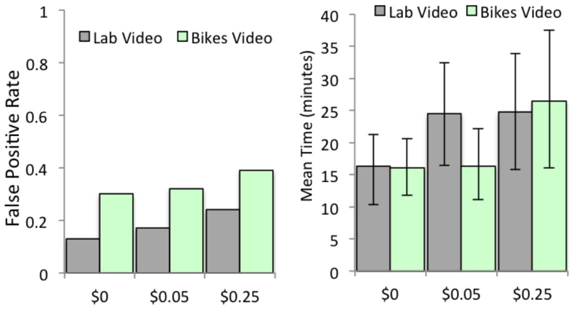


Fig. 3. Left: the overall false alarm rates calculated from users of the *live* interface, separated by video and bonus condition; right: the mean time period that the video was watched in minutes (error bars represent 95% confidence intervals).

F_1 scores in a 2 (video complexity: simple, complex) x 3 (incentive: none, low, high) ANOVA, we found a main effect of video type with workers being more accurate when watching the simple video ($F(1, 89) = 4.450, p = .038$). There was no main effect of incentive on accuracy, ($F(2, 89) = 0.909, p = 0.407$) nor was there an interaction between video and bonus ($F(2, 89) = 1.362, p = .262$).

We should note here that false alarms (clicks made to signal an alert where no relevant event was actually present) are a particular concern in a real deployment as each click must potentially be acted upon by a human i.e. there is a genuine cost-per-event identified. While the metric of precision does take into account false alarms generated by users, we also conducted an analysis for the false alarms generated in each condition. Figure 3 illustrates the false alarm rate calculated across all participants and would suggest that the false positive rate increases with the bonus. However, a 2 (Video type) x 3 (Incentive) ANOVA performed on the false alarm data revealed only a significant effect of video type ($F(1, 89) = 9.968, p = 0.002$) such that more false alarms were recorded on the bikes video.

5 Summary

In the first user study we attempted to understand how crowdsourced workers would perform when given the task of analysing live surveillance video. This is against the backdrop of all known infrastructures adopting this method of presenting video to users. We gained a number of interesting insights:

- Video type had no effect on engagement, but we had a large number of dropouts overall
- Accuracy varied with video type: workers were more accurate when watching the simple video, as predicted.
- Financial incentives had no significant effect upon either the amount of video participants were prepared to analyse, nor their accuracy.
- Increasing financial incentives superficially appeared to increase the number of false alarms generated, although this effect was not significant (see Figure 3).

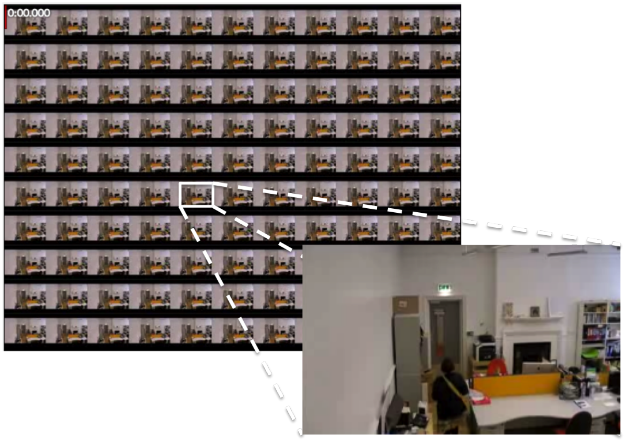


Fig. 4. Panopticon is a video surrogate system that represents an overview of the frames of a video in a grid configuration. Each tile in the grid represents a moving section of video.

6 Experiment 2 - Archived Video Surveillance

In this second study, we recruited a different set of crowd workers and asked them to discover events in fixed length sections of video; a situation representing retrospective surveillance. We introduced two different interfaces to view the same two surveillance videos we used in Experiment 1. This meant that watchers were free to navigate sections of the video and be proactive in their searches for the events we specified.

7 Method

The study consisted of a 2 (interface) x 2 (video type) x 3 (incentive level) independent design with dependent measures again reflecting engagement (time on task) and participant accuracy calculated from measures of precision, recall, and F_1.

7.1 Task Materials

Although the video footage was identical to that used in Experiment 1, for this study each video was cut into 15 minute segments (creating 8 sub-videos) with each segment searchable using one of the two experimental interfaces.

Given the novelty of Panopticon, participants assigned to that condition were guided through a fast step-by-step guide of its features and its interaction possibilities. Once the participant had clicked through the three dialogue boxes detailing the functionality of Panopticon, they were able to start the study.

7.2 Procedure

The procedure differed slightly from the previous study, as participants were firstly given an introduction to their assigned video navigation interface and then asked to use this interface to analyse as many video segments as they wished. Base payment was conditional upon them having monitored at least one 15 minute sub-video.

7.3 Participants

We recruited 359 participants through MTurk (58% male, 42% female). 181 participants took part using Panopticon while the other 178 used Scrub-Player. There were 2 dropouts for the Scrub-Player condition, and 7 from the Panopticon group (they did not start the first video).

8 Results

As with the first experiment, we analyzed the performance of the crowd through measurements of time engagement and participant accuracy. In the former, our significance testing was focused upon the number of minutes of video analysed by participants, and in the latter this was focused upon accuracy as described by an F_1 score.

8.1 Time Engaged on Task

We found a main effect of interface. Participants using Scrub-Player spent less time watching video than those using Panopticon ($F(1, 353) = 36.827, p < .001$). We also found a main effect of video type, such that participants assigned to watch the simple lab video watched for longer than those watching the complex bikes video ($F(1, 353) = 13.788, p < .001$). We found no main effect of financial incentive on the time spent analysing video, ($F(2, 353) = 0.298, p = .742$).

We found a significant interaction between interface and video ($F(1, 353) = 5.436, p = .020$) where Panopticon users spent significantly longer on task for the simple videos compared to the more complex videos and completed more tasks ($t(185) = 3.869, p < .001$). This was not true for Scrub-Player users, where no significant effect of video was found ($t(176) = 1.335, p = .184$). No significant two-way interaction effects were found between interface and bonus ($F(2, 353) = 0.048, p = .953$), video and bonus ($F(2, 353) = 1.211, p = .299$), nor was there a three way interaction between interface, video and bonus ($F(2, 353) = 0.030, p = .970$).

Figures 3 and 5 illustrate the range of time spent across each video and incentive condition and illustrates confidence intervals. In the case of workers receiving no financial incentive (for either video) they would watch on average for 16 minutes. In this study, time spent on task does not equate to the number of minutes of video that were analysed. The additional time on task translates into a higher mean number of minutes of video analysed for users of the Panopticon system, as indicated in Tables 2 and 3.

8.2 Accuracy

Again, our primary measure of accuracy was the F_1 score derived from measures of precision and recall. We found that participants using Scrub-Player were more accurate

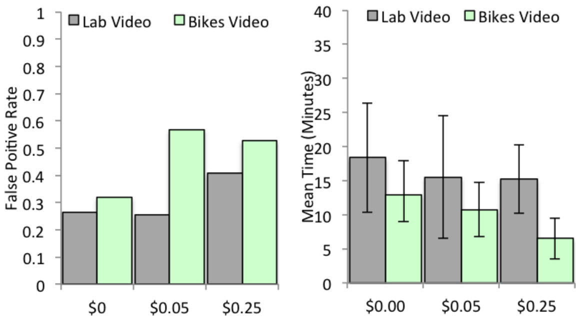


Fig. 5. Left: the overall false alarm rates calculated from users of the Panopticon interface separated by each video and bonus condition; right: the mean time period that participants spent on the task in minutes (error bars represent 95% confidence intervals).

than those using Panopticon ($F(1, 259) = 13.614, p < .001$). We also found that participants viewing the simple lab video had a significantly higher F_1 score than those viewing the complex bikes video ($F(1, 259) = 40.986, p < .001$). We found no effect of incentive on accuracy ($F(2, 259) = 0.396, p = .674$).

We were interested in exploring this data further – again looking at the issue of false alarms. We found a significant main effect of interface ($F(1, 259) = 4.031; p = .045$) with more false alarms made when using Panopticon as compared to Scrub Player; a significant main effect of video ($F(1, 259) = 21.995; p < .001$) with more false alarms made to the complex video; a marginal effect of bonus ($F(2, 259 = 2.674; p = .070$) with greater levels of pay leading to more false alarms (see Figure 5) and a marginal interaction between video and bonus ($F(2, 353) = 2.767; p = .064$), such that the performance 'cost' of paying high bonus rates is most apparent for workers engaged in the simple lab task.

9 Summary

In this second study we modelled the task of workers who must search archived footage and can, consequently, employ video search tools to improve efficiency. We introduced two such tools or interfaces: the first was a standard Scrub-Player that allows rapid fast-forward and backward scanning in a video task; the second, Panopticon was a new system, shown to be more effective than a scrub player in a series of video search tasks [15]. We had predicted that Panopticon would be more effective in supporting video surveillance searching, but our findings were more nuanced – Panopticon users watched more footage, with the result that they found more events, but accuracy was poorer than with the standard Scrub-Player and workers made more false alarms, particularly when watching the complex video when event detection was relatively poor. We found no overall effect of financial incentive, however there was some indication in the complex video case that bonus payments can simply lead to the production of more false alarms, particularly when workers are presented with uneventful video.

Table 2. Descriptive statistics of the accuracy of participants using *Panopticon* to watch: (top) the lab video and (bottom) the bikes video.

Bonus	n	Mins of Video (μ)	Precision (μ)	Recall (μ)	F1 (μ)
$0	29	67.5	0.73	0.72	0.69
$0.05	29	57	0.72	0.64	0.67
$0.25	31	60	0.67	0.64	0.66
Total	89	61.5	0.71	0.67	0.67

Bonus	n	Mins of Video (μ)	Precision (μ)	Recall (μ)	F1 (μ)
$0	33	40.5	0.53	0.37	0.46
$0.05	30	42	0.56	0.4	0.47
$0.25	29	42	0.33	0.28	0.26
Total	92	41.5	0.47	0.35	0.4

10 Discussion

There is a perception that streaming surveillance video to online platforms is a lazy method to obtain security that can only serve to harden the prejudices that exist in society. For instance, it is already known that in the traditional delivery of CCTV video to the control room that the attention of watchers does not fall equally upon all people [25]. Ethnographic studies of CCTV control rooms have suggested that cameras are particularly drawn to minority groups, such as homeless people; archetypes of Bauman's failed consumers [1]. While the discussion of privacy must happen in more detail elsewhere, our endeavour was to study the performance parameters of opportunistically recruited members of the crowd on the task of monitoring surveillance video. To these ends, we have carried out two studies assessing the feasibility of using a crowdsourcing platform to detect events in CCTV surveillance video. We presented both live and archived video footage and investigated two factors likely to impact the performance of the crowd: interface type and incentive scheme in order to learn more about the efficacy of using the crowd for video surveillance work. Our discussion reflects the insights generated from these empirical studies and includes suggestions for future work.

10.1 Live vs Archived Video

The live video stream in isolation appeared to poorly engage the crowd workers. A large number of participants did not complete the basic unit of work that we specified; possibly due to the open-ended nature of the task, i.e. there was no foreseeable end to the task, nor were there any significant milestones. We already know that those employed to monitor live video feeds find this task onerous and experience high levels of boredom [25], and so we would argue that where circumstances allow, a retrospective search of archived material is preferable. However, where there remains a need for watchers to monitor video surveillance in ‘real time’, it might be useful to develop

Table 3. Descriptive statistics of the accuracy of participants using the *Scrub-Player* to watch: (top) the lab video; (bottom) the bikes video.

Bonus	n	Mins of Video	Precision	Recall	F1
		(μ)	(μ)	(μ)	(μ)
$0	31	37.5	0.83	0.69	0.73
$0.05	29	27	0.81	0.64	0.79
$0.25	28	34.5	0.85	0.75	0.84
Total	88	33	0.83	0.68	0.79

Bonus	n	Mins of Video	Precision	Recall	F1
		(μ)	(μ)	(μ)	(μ)
$0	30	25.5	0.7	0.69	0.63
$0.05	32	30	0.72	0.6	0.61
$0.25	28	28.5	0.46	0.56	0.5
Total	90	28.5	0.63	0.62	0.43

efficient algorithms that separate the crowd into those monitoring real-time footage, and those conducting retrospective analysis to double-check alarms raised (a technique quite commonly used in citizen science [24]).

10.2 Interface Design

We noted that participants using Panopticon analysed significantly more video than those using the Scrub-Player. They also raised more correct alarms in total, but were less accurate overall due to the tendency to raise more false alarms as well. We believe the relative visual complexity of the Panopticon interface (particularly given workers' lack of familiarity with that system) contributed to a higher rate of false alarms and lower levels of accuracy. The Scrub-Player provided a simpler, more familiar interface and resulted in a significantly higher F_1 (accuracy) score, but this interface does require intricate motor skills. Therefore, designers should consider both the costs and benefits of increasing the complexity of user interfaces applied to the task. Image processing techniques could also be used to support the selection of segments of footage for further analysis by a bespoke interface [6, 19].

10.3 Task Complexity

We noted a significant effect of the video type in both studies which suggests that the characteristics of the surveillance video can significantly impact the quality of work that can be expected from the crowd. This affects our consideration of the most appropriate user interfaces to support crowd work. In tasks monitoring both live and archived video footage, accuracy dropped for the more complex video. We received messages, generated by watchers of the bikes video that showed they were trying to indicate ways in which other events might have been of interest e.g. *"The truck is in the way of the*

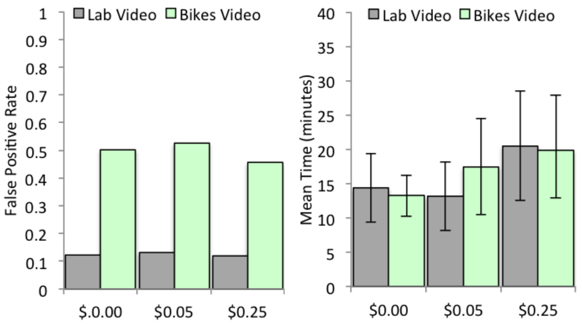


Fig. 6. Left: the overall false alarm rates calculated from watchers of each video and bonus condition for the Scrub-Player; right: the mean time period that participants spent on the task (error bars represent 95% confidence intervals).

camera!", *"...those builders are just hanging around, they need to get back to work"*. However, none of the interfaces we provided were sufficiently expressive to enable such concerns to be distinguished from critical events. For complex videos, facilitating different types of alerts could create a more engaging eco-system for crowd workers, so that their efforts still feel valued even if such extra information is not considered necessary. It may, however, help to disaggregate the valuable from non-valuable alarms, particularly in videos that could generate a prohibitive number of false alarms.

10.4 Financial Incentives

We noted a limited impact of the per-event bonus offered to participants across both user studies, and we even found marginal effects where higher bonus payments led to lower accuracy. Initially, it seems surprising that bonuses failed to incentivise participants on a platform like Mechanical Turk where the main focus of the participant is thought to be on earning rather than engaging [12]. This also runs contrary to literature that shows financial incentives can lead to better performance in resume rating [11] and web searching [16] scenarios. However, our results fit more closely with other work [20] that has suggested that payment only improves the quantity of responses from the crowd, without any benefit to the quality of work returned. Indeed some studies involving real world scenarios have even shown where money was offered as reward for a normally volunteer-based activity, paying for the volunteers time actually reduced the number of hours on average a participant contributed [9]. Of course, it might just be that our bonuses were not sufficiently attractive to influence participation or accuracy. Overall, our findings suggest that, in this surveillance context, there is a delicate balance to be considered between intrinsic and extrinsic motivation.

10.5 Identifying Genuine 'Alarms'

An inflated false alarm rate is a debilitating matter in a world of online CCTV platforms; without a sensible way to identify genuine alarms, each signalled 'event' needs to be

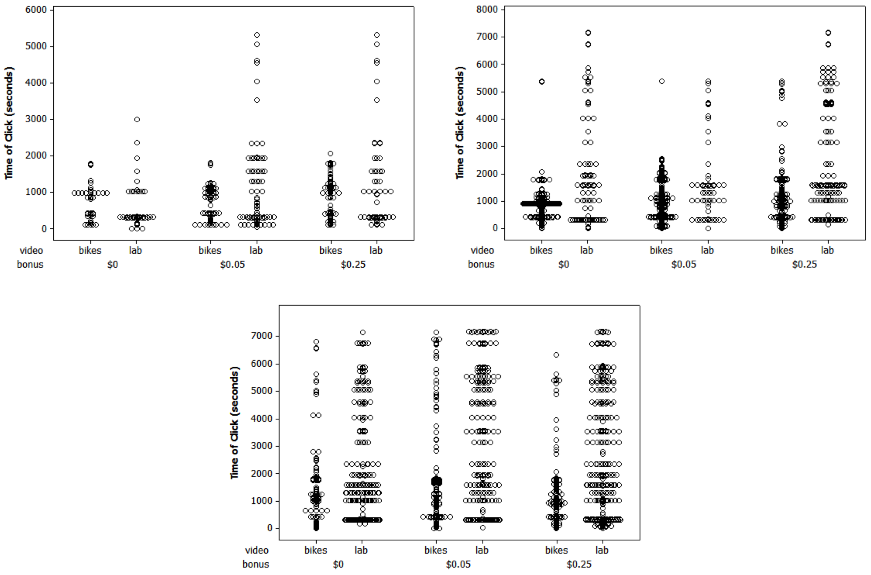


Fig. 7. The overall distribution of alerts raised by experiment participants using: (i) live interface; (ii) Scrub-Player; (iii) Panopticon. Each graph is split by bonus scheme and video type. Each circle indicates an alert raised during the 2 hour surveillance video. A greater clustering of circles indicates higher concentration of alerts raised during that time period (7200 seconds = 2 hours).

propagated up the chain of command. In our experiments we already knew where in the video the genuine events were to be found. This begs a question of whether crowdsourced analysis should be done at the level of the individual worker, or whether a group of workers should be tasked to search the same video, in which case the crowd working in concert should create reliable spikes of activity when genuine events are present. Figure 7 illustrates the click streams collected across both studies for the same videos. We could imagine that some activity threshold could be applied to data streams like this, to determine whether a particular section of video should be investigated further. Of course other methods known in the crowdsourcing community are relevant here, including traditional methods where one set of workers tags events and a second set of workers votes on those tags. Such developments are hybrid systems that could potentially cope with both live and non-live analysis, as we noted earlier, but would raise important questions about how to build trust in the work that is collected. One such hybrid system for video analysis was proposed by Velastin [28] where image processing is proposed as a first layer of processing to be supplemented by human judgement in cases that require further action.

11 Study Limitations

As with any research there are a number of limitations to our empirical work. The overall context that we attempted to model was that of a real platform asking participants to analyse real CCTV footage. As our videos did not contain any actual crime, any sense of intrinsic motivation felt by our participants may have been short lived, as such, our results provide the most reliable insight into the role of extrinsic motivators. Also, despite the lack of crime in the videos, the underlying strategies that participants would employ to spot events are likely to be similar, giving us confidence that we did model realistic video search behaviour under our different conditions. Future community-based studies could provide more reliable insight into the role of intrinsic motivators on such platforms. While the levels of activity within CCTV feeds can be diverse depending on the context, our videos represented a best-case scenario for the participants as events of interest were relatively frequent, this means that for video feeds where events are less frequent (or non-existent), designers could treat our results as an upper bound for accuracy or participation. Design decisions that we made on our platform could also have impacted the alarm raising behaviour of participants, for example, participants were not penalised for generating false alarms which may have led to a more liberal approach to raising them.

12 Conclusion

The recent rise in the number of web platforms that publish CCTV video should sharpen our focus as a research community as to whether this approach is a lazy form of security, or whether citizen participation can indeed yield security benefits. In this paper we took the first steps in this debate. We conducted two user studies on MTurk to explore the impact of financial bonus, surveillance video complexity, and video navigation interface upon the ability of the crowd to analyse surveillance video. We discovered that, across our two experiments, considerations of video complexity were most important; higher bonus payments did not encourage higher accuracy and that, for retrospective surveillance, matching video search interfaces to the video type is crucial to influence performance. We also noted that designers must be clear about the nature of the participation that is desired from the crowd, particularly with regard to whether a precision or recall focused strategy should be prioritised for event detection. Finally, we suggest there is a need to better understand the security and privacy experiences [7] that may accompany wider deployment of these infrastructures. There is an opportunity for platforms of this type to act as a democratising force to the imposition of visual surveillance in our societies, but future research is needed to determine whether future platforms can live up to this challenge, or will simply serve to further erode privacy.

13 Acknowledgments

This work was funded by Research Councils UK on the Digital Economy theme, via the Social Inclusion through the Digital Economy (SiDE) Research Hub (EP/G066019/1).

References

1. Zygmunt Bauman. *Consuming Life*. Polity Press, 2007.
2. Michael S. Bernstein, Greg , Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: A word processor with a crowd inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 313–322, New York, NY, USA, 2010. ACM.
3. Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 33–42, New York, NY, USA, 2011. ACM.
4. Live texas border watch. `http://www.blueservo.net`. Accessed: 17/09/2013.
5. Matthew JC Crump, John V McDonnell, and Todd M Gureckis. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410, 2013.
6. Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(10):1337–1342, 2003.
7. Paul Dunphy, John Vines, Lizzie Coles-Kemp, Rachel Clarke, Vasilis Vlachokyriakos, Peter Wright, John McCarthy, and Patrick Olivier. Understanding the experience-centeredness of privacy and security technologies. In *Proceedings of the 2014 Workshop on New Security Paradigms Workshop*, NSPW '14, pages 83–94, New York, NY, USA, 2014. ACM.
8. Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, CHItaly '13, pages 14:1–14:4, New York, NY, USA, 2013. ACM.
9. Bruno S Frey and Lorenz Goette. *Does pay motivate volunteers?* Institute for Empirical Research in Economics, University of Zurich, 1999.
10. B Grissom. Border watch meets lowered expectations in fourth year. `http://www.elpasotimes.com/ci_14918540`. Accessed: 17/09/2013.
11. Christopher Harris. Youre hired! an examination of crowdsourcing incentive models in human resource tasks. In *WSDM Workshop on Crowdsourcing for Search and Data Mining (CSDM)*, pages 15–18, 2011.
12. John Joseph Horton and Lydia B. Chilton. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, EC '10, pages 209–218, New York, NY, USA, 2010. ACM.
13. Internet eyes. `http://www.interneteyes.co.uk`. Accessed: 17/09/2013.
14. Dan Jackson, James Nicholson, Gerrit Stoeckigt, Rebecca Wrobel, Anja Thieme, and Patrick Olivier. Panopticon: A parallel video overview system. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 123–130, New York, NY, USA, 2013. ACM.
15. Dan Jackson, James Nicholson, Gerrit Stoeckigt, Rebecca Wrobel, Anja Thieme, and Patrick Olivier. Panopticon: a parallel video overview system. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 123–130. ACM, 2013.
16. Gabriella Kazai. In search of quality in crowdsourcing for search engine evaluation. In *Advances in information retrieval*, pages 165–176. Springer, 2011.
17. Walter S. Lasecki, Christopher D. Miller, and Jeffrey P. Bigham. Warping time for more effective real-time crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2033–2036, New York, NY, USA, 2013. ACM.

18. John Lettice. London estate broadband offers 'spot the asbo suspect' tv channel. `http://www.theregister.co.uk/2005/12/30/shoreditch_digital_bridge/`. Accessed: 12/02/2014.
19. Suzanne Little, Iveel Jargalsaikhan, Kathy Clawson, Marcos Nieto, Hao Li, Cem Direkoglu, Noel E. O'Connor, Alan F. Smeaton, Bryan Scotney, Hui Wang, and Jun Liu. An information retrieval approach to identifying infrequent events in surveillance video. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, ICMR '13, pages 223–230, New York, NY, USA, 2013. ACM.
20. Winter Mason and Duncan J Watts. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.
21. Michael McCahill and Clive Norris. Cctv in london. *Report deliverable of UrbanEye project*, 2002.
22. Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowd-sourcing markets. In *ICWSM*, 2011.
23. Aaron D. Shaw, John J. Horton, and Daniel L. Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 275–284, New York, NY, USA, 2011. ACM.
24. S. Andrew Sheppard and Loren Terveen. Quality is a verb: The operationalization of data quality in a citizen science community. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 29–38, New York, NY, USA, 2011. ACM.
25. Gavin JD Smith. Behind the screens: Examining constructions of deviance and informal practices among cctv control room operators in the uk. *Surveillance & Society*, 2(2/3), 2002.
26. Gavin JD Smith. Exploring relations between watchers and watched in control (led) systems: strategies and tactics. *Surveillance & Society*, 4(4), 2002.
27. Doug Tewksbury. Crowdsourcing homeland security: The texas virtual borderwatch and participatory citizenship. *Surveillance & Society*, 10(3/4):249–262, 2012.
28. SergioA. Velastin. Cctv video analytics: Recent advances and limitations. In Halimah Badioze Zaman, Peter Robinson, Maria Petrou, Patrick Olivier, Heiko Schrder, and TimothyK. Shih, editors, *Visual Informatics: Bridging Research and Practice*, volume 5857 of *Lecture Notes in Computer Science*, pages 22–34. Springer Berlin Heidelberg, 2009.
29. Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.
30. Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *Int. J. Comput. Vision*, 101(1):184–204, January 2013.
31. Carl Vondrick, Deva Ramanan, and Donald Patterson. Efficiently scaling up video annotation with crowdsourced marketplaces. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 610–623, Berlin, Heidelberg, 2010. Springer-Verlag.