Security Journal (2021) 34:389–409
https://doi.org/10.1057/s41284-020-00230-w

ORIGINAL ARTICLE

Automated monitoring for security camera networks: promise from computer vision labs

Chen Chen[1] · **Ray Surette**[2] · **Mubarak Shah**[3]

Published online: 17 February 2020

Abstract

A substantial increase in the number of surveillance camera systems has not delivered the promised deterrent effects or investigative case evidence and their usefulness has been underwhelming. A potential solution to practical camera monitor needs is computer vision (CV)-enhanced camera networks that can provide automated real-time video analysis, quick processing of monitor query-based searches, and accurate summaries of archived video files. The development and testing of four CV algorithms in computer vision laboratories is presented and implications from their possible adoption by security agencies on society are discussed.

Keywords Computer vision · Crime prevention · Camera networks · Security cameras · Camera monitoring · Surveillance cameras

Introduction

Driven by the availability of less expensive cameras, contemporary surveillance has shifted from a human-based activity to the one dominated by camera technology. Despite long-standing concerns regarding the effects of camera surveillance on society, cameras remain a popular choice to address various security concerns and today cameras are common in various public and private sites (Adams and Ferryman 2015; La Vigne et al. 2011; Sandhu 2017; Scheitle and Halligan 2018; Surette 2015). Regarding their effectiveness, reviews of police surveillance cameras have reported human-monitored surveillance cameras to be effective in some settings for some crimes, but consistent positive impacts have not been found (Alexandrie 2017;

✉ Ray Surette
Raymond.surette@ucf.edu

[1] Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

[2] Department of Criminal Justice, University of Central Florida, Orlando, FL 32816-2365, USA

[3] Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816-2365, USA

Ashby 2017; Gerell 2016; Gill 2003; Goold 2004; La Vigne et al. 2011; Piza et al. 2019; Prenzler and Wilson 2019; Taylor 2010; Welsh and Farrington 2002, 2004; Welsh et al. 2015). In a recent meta-analysis, Piza et al. (2019) found that public space camera networks were associated with a modest but significant decrease in crime, the largest and most consistent being observed in car parks.

Most of the extant literature on surveillance cameras, however, has examined their proactive value for crime prevention and crime reduction. There is comparatively little research on reactive applications (Ashby 2017). Beyond crime deterrence, a surveillance camera application described as having great potential but not heavily examined is the use of camera footage for investigations. The lack of research attention is despite the long-standing argument that surveillance camera networks work better for investigations than for crime reduction. Honovich (2008), for example, stated that surveillance camera systems were best suited for crime solving rather than crime reduction and should be used in crimes where offenders conduct pre-crime risk assessments. Noting that in the private sector the use of surveillance cameras is overwhelming, which are justified as crime investigation tools, Honovich (2008) argued that camera network operators should focus on solving rather than deterring crime. Despite this early call, subsequent research focused on studying crime prevention effects and research on the use of surveillance cameras for investigative purposes remained limited in spite of the fact that when available police regularly request surveillance video when conducting investigations (Morgan and Coughlan 2018).

The prior research on surveillance cameras' value for investigations that is available is also often not rigorous. A substantial amount is journalist based and published as news stories (cited by Ashby 2017 see for example Davenport 2007; The Scotsman 2008; Bulwa and Stannard 2007, and Edwards 2008, 2009). Surveillance cameras might be useful for criminal investigations because they can directly help answer two basic investigative questions: what happened, and who was involved (La Vigne et al. 2011). Useful video increases the initial detection of crimes and provides reviewable evidence regarding what and who and footage can allow investigators to watch an entire incident and corroborate or refute other evidence or testimony. An arrest need not result for footage to be useful and the elimination of a suspect or a crime is also socially beneficial. In a recent Australian-based study, Morgan and Dowling (2019) found a 20% increase in clearance rates when requested video was delivered.

Under what circumstances are surveillance cameras most likely to be useful? Older reports of how police view the usefulness of surveillance camera footage are mixed. Some reported that police see it as highly valuable, others as counterproductive—to the point that Ashby (2017) notes that some have recommended the cessation of human monitoring. Surveillance camera video does appear useful though for increasing detection for many crimes, particularly for robbery and violent crimes (Ashby 2017). According to the survey of police investigators in Australia, 9 of 10 officers highly valued camera footage (Dowling et al. 2019). Three of four surveyed investigators felt footage when available was useful or very useful. They further perceived camera footage as particularly useful in the earlier stages of investigations and liked camera footage for investigating assaults while acknowledging that in

their view surveillance video best increased clearance rates for theft, burglary, and property damage (Dowling et al. 2019). The most common uses mentioned were to identify subjects, followed by developing leads, corroborating witness, and suspect statements, and determining if a crime had occurred. Overall, the extant research strongly suggests that surveillance cameras can be powerful investigative tools for many types of crime.

Irrespective of their popularity, there remains a substantial gap between what was promised and what has been delivered by camera networks (Prenzler and Wilson 2019). One reason for the gap is because the number of cameras in many networks outpace human capacity to effectively monitor them (Donald 2005; Hesse 2002; Prenzler and Wilson 2019; Welsh et al. 2015). In practice, camera monitors usually have two tasks: general monitoring of multiple live camera feeds or searching for a specific event, person, or object in archived video files, usually associated with an investigation. However, even when vigilant, human monitors quickly become cognitively swamped, frequently missing important content (Faber et al. 2012; Sasse 2010). Contemporary camera systems are therefore hit-or-miss tools regarding the observation and detection of ongoing incidents and expensive time-consuming search platforms for finding specific video sequences (Gill 2003; Goold 2004; Hier et al. 2007; Näsholm et al. 2014; Ratcliffe et al. 2009; Sasse 2010).

An increasingly popular approach to the shortfalls of human-monitored camera networks is computer vision (CV). In addition to practical improvements in processing video, CV has the potential to address problems associated with inappropriate use of surveillance for profiling and voyeurism and from unintended human monitor errors due to inattentional blindness and attention capture failures.[1] The promise of CV-enhanced cameras is the reduction of the two main human sources of ineffectiveness regarding security cameras. A computer algorithm will not become bored or distracted during real-time monitoring and events of interest are more likely to be quickly found (Idrees et al. 2018). Research on the effectiveness of CV software to automatically analyze large camera networks has been ongoing (Adams and Ferryman 2015; Coetzer et al. 2011; Gong et al. 2011; Gowsikhaa et al. 2014; Hesse 2002) and calls for CV's incorporation into surveillance camera systems and discussions of potential applications began to appear in the early 2000s (see Baldwin and Baird 2001; Barrett et al. 2005; and Thomas and Cook 2006). CV's full potential has not been exploited however and contemporary applications are concentrated in facial recognition and license plate readers (Adams and Ferryman 2015). The purchase of camera systems that require human monitors continues to be the common practice (Keval and Sasse 2010; Piza et al. 2014a).

Countering this trend, the recent coupling of security cameras to inexpensive computers equipped with fast graphic processing cards (GPUs) has put sophisticated

[1] Monitor perception failure occurs when there is little visual change present in long video stream stretches and a monitor's attention shifts to non-visual tasks such as conversations or daydreaming (Bredemeier and Simons 2012; Fougnie and Marois 2007; Mack and Rock 1998; Memmert 2006; Most et al. 2005; Sasse 2010). Monitoring video streams has been reported to significantly increase perceptual failure (Hyman et al. 2009; Most et al. 2005).

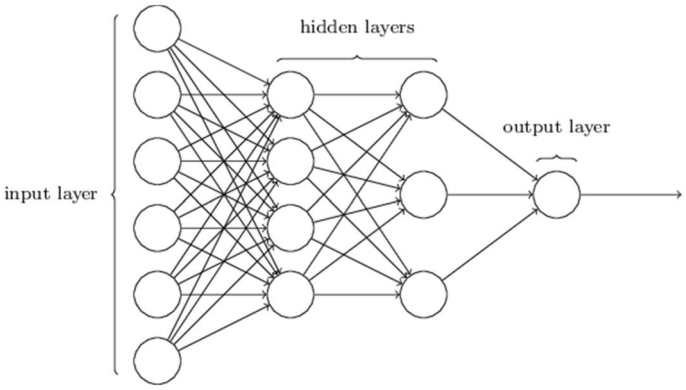


Fig. 1 A simplified neural network consisting of input (images) and output (labels) layers, and two intermediate hidden layers. The circles represent the linked nodes. Each link has a weighted parameter associated with it that is "learned" during training (See Abdi et al. (1999) for an introduction to neural networks geared for social scientists)

CV applications within reach of security firms and public agencies. In particular, the development of artificial neural networks (see Fig. 1) which consist of a vast number of linked quantitative "neurons" nodes in layered setups represent a significant advancement in CV science. In a typical CV application, the input to the neural network consists of training images, and output consists of semantic labels corresponding to those images. The network quantifies each intervening layer automatically so that a neural network "learns" to correctly connect the output labels to the input images. The training of a network can involve millions of numerical parameters that link hundreds of thousands of multi-layered neural nodes. Large neural networks have been found to significantly improve performance on various computer vision tasks such as object detection, face identification, and action recognition while reducing computation time (Adams and Ferryman 2015). However, large neural networks require large amounts of training data compared to traditional machine learning algorithms and thus may be limited for some security applications.

Although readily available, visual data generated by security cameras have not been exploited due to inherent difficulties in processing digital images (Idrees et al. 2018). Until recently, visual data files were simply too massive and CV hardware and software were too slow to be useful. As computational capabilities and processing speed improved, an accompanying set of sophisticated CV applications have begun to appear. Digitized video is being seriously examined as both a reactive post-event data source and as a means to generate proactive pre-event predictions (Adams and Ferryman 2015; Ferguson 2017). Irrespective of ongoing developments in CV research and long-standing calls for automated analysis (see for example, Hesse 2002), discussion of recent developments in CV-based solutions to security tasks has not been substantial. The literature is dominated by reviews of license plate readers and facial recognition programs and the evaluation of human-monitored camera networks and there is a gap in the extant literature regarding developments in CV security camera analysis.

Results from Lab tests of four CV security applications

While acknowledging the inherent limitations related to the lab testing of technology in lieu of field tests, as an initial evaluation step, this article describes and presents computer vision lab tests for four contemporary CV applications that address core camera network monitoring tasks. Each application is herein examined, their CV lab test results are summarized, and the implications of their real-world use are discussed. The four CV applications tested are the following: the real-time detection and labeling of objects and events, the detection of anomalies, the automatic summation of long videos, and the search of long video files based on human queries. While simplified versions of these tasks have been explored in computer vision science using clean, unambiguous videos, they have not been tested from a real-world perspective using security camera videos. Instead, prior research has focused on actions and events of little interest to security professionals (see for example Yang et al. 2009). And while CV research has reported high algorithm accuracy for comparatively easy data sets, accuracy had not reached levels adequate for real-world applications. For example, it has been found that action recognition performance is above 90% for easy datasets and simple actions such as running. But accuracy drops to around 60% for challenging real-world surveillance video datasets and actions such as fighting and theft—an accuracy rate which would correlate with a substantial number of false reports from security camera networks (Kuehne et al. 2011). As raw in-the-wild videos, security camera videos remain a challenge for CV algorithms. Regarding the CV lab assessments, it should be noted that computer vision science does not assess algorithms by applying statistical significance tests commonly used in the social sciences. Instead, CV programs are evaluated either in comparison with prior CV methods on standard visual image data sets or compared against the performance of humans (or 'ground truth' comparisons).

Task 1: action and event detection

Applicable to both live and archived video files, the detection and labeling of actions and events in surveillance videos is a basic monitoring need that is easily done by humans, but has not been consistently achieved in CV (see for example Yang et al. 2009). Example objects of interest would include weapons, unauthorized cars, and abandoned property. Criminal events of interest that potentially could be detected by CV would include graffiti, vandalism, theft, robbery, and batteries; non-criminal events would include car crashes, crowd stampedes, injuries, fires, and explosions. In CV algorithms, an action or event is usually conceived as a sequence of sub-sections in a specific order. For example, a robbery can be decomposed into a person A approaching person B, person A producing a weapon, gesturing at person B, person B holding their hands aloft and surrendering their property, and the two separating. Two basic CV concepts are helpful for understanding CV-labeling objects and events. The first is a 'classifier'—a computer vision sub-routine that recognizes and assigns labels to objects and activities in images. Classifiers can answer queries about unlabeled

Fig. 2 Positive training examples for assault, theft, graffiti, and robbery

images such as Is there a handgun in this video file? The creation of classifiers in turn invokes the second CV concept 'training.' CV classifiers require 'training' to empirically improve their accuracy. The common process for classifier training on a previously unlabeled object starts with a human providing hundreds or thousands of representative images of an object or activity of interest. Training normally requires two types of examples– a set of correct images that show the in-subject variation of an object (i.e., how guns can be visually different from each other, but still be 'guns') and a set of incorrect examples of things that are not guns, but might be mistaken for one (a spray bottle can have a trigger for example). Provided with enough positive and negative examples, training allows a CV program to mathematically differentiate objects and evaluate unlabeled visuals. An approach that utilized an extension of CV deep learning neural networks termed tube convolutional neural network or T-CNN was developed and tested for the task of spatial localization of objects and recognition of events of interest in security videos (Shah 2017) (Fig. 2).

CV example: a tube convolutional neural network

This CV method for detection of static objects and dynamic events employed two previously established CV functions: semantic indexing (applying a label like "robbery" to a video sequence or "gun" to an array of pixels in a single frame) to label objects such as weapons, security personnel, and official vehicles, and event detection for determining dynamic complex actions like assaults, thefts, crashes, and explosions. Previously, the impact of deep learning on video analysis for tasks such as action recognition had been limited due to the inherent complexity of video data and the limited availability of annotated training videos. An advance of T-CNN event detection was that it evaluated entire videos and categorized them into broad

Fig. 3 An example of action (fighting) detection results for T-CNN. The green boxes indicate the ground truth and the red boxes indicate the predicted bounding boxes for action "fighting" generated by CV. The numbers denote the probabilities associated with a correct label being assigned to the action

classes of video clips rather than simply cataloging brief appearances of single objects. The aim of this detection and classification system was to better handle more challenging real-world video analysis and to accurately classify video segments into useful actions such as car crashes, robberies, and assaults.

Regarding action detection, previous deep learning-based action detection approaches worked by first detecting potential single frame-level acts associated with pre-labeled common actions (i.e., a single image of someone running or jumping was identified) or alternatively produced action labels following extensive time-consuming CV classifier training (Uijlings et al. 2013). These prior approaches did not use the temporal information available in video files however. In T-CNN, the simultaneous analysis of time and location information was exploited and the addition of "time" information allowed the T-CNN program to recognize and localize actions based on both spatial (location in a frame) and temporal (flow across frames) image data. Thus, an action such as fighting would be located in time in a video sequence by automatically determining the frames where the action began and ended and localized spatially to the pixels in each frame associated with fighting. A lab test of the T-CNN action detection method used a data set in which all videos were captured in realistic settings and included intra-class variation, camera motion, differing viewpoints, and scale changes. The data set was split by taking one third of the videos from each action category to form a test set, with the balance used for algorithm training. A standard CV assessment – receiver operating characteristic (ROC) curve comparison – was applied and the results for T-CNN compared favorably to two competing approaches.[2] An action detection example is shown in Fig. 3.

[2] The first alternate approach (termed tubelets) utilized selective sampling to produce sequences of bounding boxes for action localization (Jain et al. 2014). The second competing approach, termed poselets, developed a relational model for action detection which initially decomposes human actions into temporal 'key poses' and then into spatial 'action parts' (Wang et al. 2014). Our method (T-CNN) was compared with these two prior state-of-the-art action detection approaches on our collected real-world action detection dataset including crime-related events/actions such as fighting, car accident, and robbery. The ROC curves of these approaches were plotted. At each False-Positive Rate, the higher the "True Positive Rate" the more accurately the method detects actions. From the results, our approach was superior to the two alternate CV approaches. However, our method sometimes also missed the real events. For example, when the event/action region is very small in the video due to long distance camera view, our method missed the detection due to limited information. False positives occur when events/actions are very similar in terms of appearance or motion like robbery and burglary. This would confuse the CV algorithm, leading to false positives for those actions such as labeling a robbery as a burglary. The figure below reflects that the T-CNN approach was superior to the two alternate CV approaches.

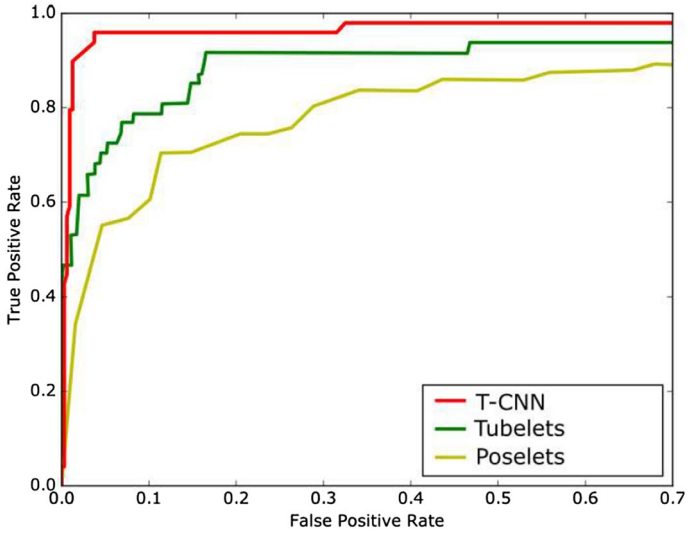
Task 2: Anomaly detection

Of particular value to security camera monitors are activities which occur infrequently, but are precursors to criminal activity (for example, 'car hopping' where a person pulls on car door handles as they walk along a street would be a precursor to theft from vehicles). These activities are more difficult to program in CV simply because they are rare and examples needed for classifier training are often unavailable. The solution is a CV program with the ability to flag anomalous, previously unseen events. Adequate anomaly detection algorithms must continuously update and incorporate slow, subtle changes to their visual environments because detection of abnormal behaviors has been an inherently difficult CV task. The CV problem amounts to consistently flagging important new patterns in streaming digital image data that significantly deviate from historical benign patterns. First, a quantitative definition of a normal visual and abnormal visual pattern is not well defined. Second, normal behavior evolves over time and may change significantly. For instance, many people walking during daylight versus few people walking at night visually differ, but both are normal for downtown urban centers and anomalous if reversed.

CV example: detecting anomalies using a weakly supervised search approach

An example CV solution for abnormality detection begins with training videos preidentified as containing anomalies in a 'weakly supervised' training approach (Zhou

Footnote 2 (continued)



A few real-world action/event detection examples using the computer vision algorithm can be viewed from here: https://docs.google.com/presentation/d/1MINyHYIuotHTttUrjSKdCIKuR_LrW4eNCht_0kiDjgU/edit?usp=sharing.

2018). "Weakly supervised" means that each training video is labeled as containing an anomaly or not containing an anomaly, but the location of the anomaly in the video is unknown. The associated security application would be searching a long 48-h video known to somewhere contain a crime. Previously, sparse coding-based approaches represented the state-of-the-art CV anomaly detection methods. These methods optimistically assume that only the small initial portion of a video contains normal events, and therefore, the initial video portion was used to build a "normal event" dictionary. In these approaches, anomalies were empirically defined as events that cannot be re-constructed from the normal event dictionary; their quantitative uniqueness in comparison to the rest of the video file identifies them as anomalous. Although such unsupervised approaches were appealing, they were based on the questionable assumption that any pattern which deviates from the initial visual patterns in a video is an anomaly. As it was difficult to accurately define the "normal" region of a video in a way that took all possible normal patterns into account, CV-based anomaly detection proved ineffective.

With these caveats in mind, an anomaly detection algorithm using weakly labeled training videos was created and tested where videos were pre-labeled as normal or anomalous. To conduct the weakly supervised learning, a "multiple instance learning" (MIL) approach was utilized (Andrews et al. 2003). In MIL, the CV algorithm receives a set of labeled video samples (termed "bags" in CV science) that have been pre-identified by a human as containing an anomaly. Each video bag, in turn, contains many unlabeled instances obtained from the original long videos that have been divided into short segments. Specifically, the identification and location of anomalies is achieved by treating normal and anomalous surveillance videos as composed of multiple examples of activities that are possibly normal or abnormal. [3]This approach allowed for rapid CV algorithm training. The CV lab development employed surveillance videos from YouTube using queries such as car accident, armed robbery, bar fight, and theft across both indoors/outdoors and day/night scenes resulting in 940 anomalous and 117 normal training videos for analysis. For algorithm testing, 70 long videos from YouTube and surveillance cameras from a municipal police department were used reflecting different locations, weather conditions, and types of anomalies. Examples from the training and testing videos are shown in Fig. 4.

The accuracy of the 'weakly supervised' approach was tested by comparison against human ground truth results (examples results are shown in Fig. 5). In this figure, the first and second rows show anomalous and normal videos, respectively. Figure 5a involves arson, and the shaded box indicates the ground truth for the anomaly. As reflected, the anomaly detection approach generated high anomaly scores (close to 1). Similarly, Fig. 5b shows a correct detection of a road

[3] The method generated a regression model such that anomalous video segment instances have higher anomaly R^2 scores than the normal segments. The anomaly scores are not identical to the R^2 values familiar to social science research, but they analogously vary between 0 and 1 with scores closer to one denoting more anomalous video clips. To manage output levels and avoid event swamping, threshold values can be chosen to decrease (closer to zero) or increase (closer to 1) the number of clips labeled as anomalies.

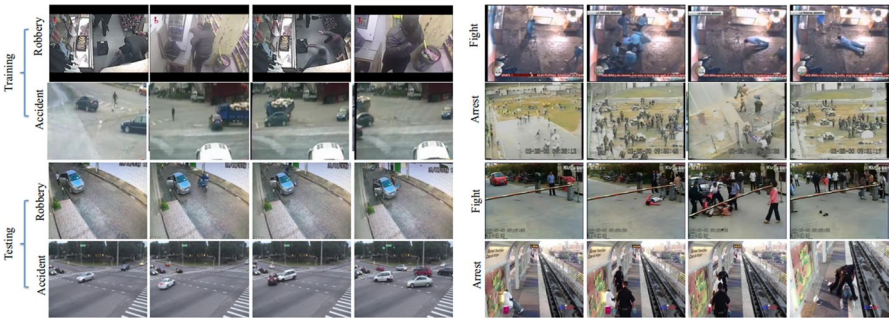


Fig. 4 Examples of different anomalies from the training and testing videos. The comparison anomaly datasets contained fewer and shorter videos of simple anomalies (e.g., the appearance of bikers, running, moving in the opposite direction). The test dataset consisted of more realistic anomalous events with substantial variation in scenes, viewpoints, illuminations, and other visual factors and concentrated on anomalies that would be of interest to police agencies such as fighting, accidents, robbery, murder, vandalism, thefts, and arrests

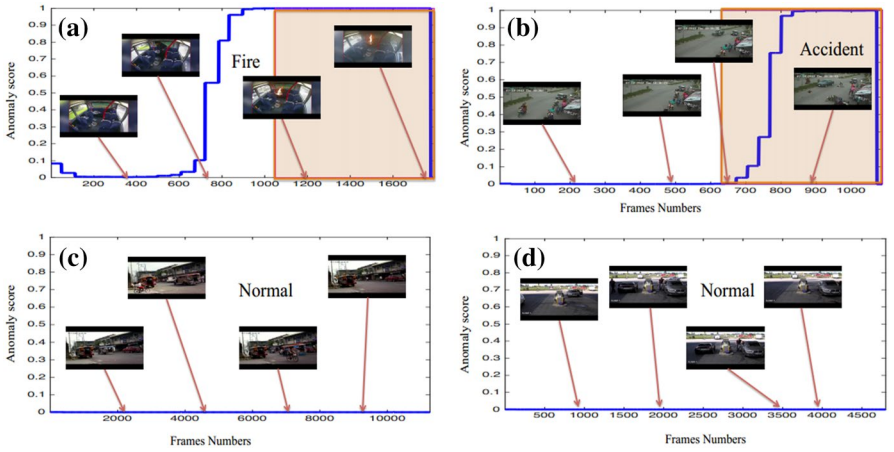


Fig. 5 Results of weakly supervised anomaly detection. Shaded window shows ground truth anomalous regions, **a** containing 'fire' and **b** 'accident' and **c**, **d** showing normal videos containing no anomaly. Weakly supervised anomaly detection algorithm was able to successfully detect anomaly regions in (**a**, **b**) by predicting high anomaly regression scores for the video frames in those regions, while also predicting low near-zero anomaly scores for normal video frames in (**c**, **d**)

accident in a video of a complex outdoor scene. Figure 5c, d report the regression scores from normal videos. Since the two videos in (c) and (d) only contain normal events, the anomaly scores were correctly low (close to 0) for both, a desirable outcome since most security videos will contain only normal events and a low alert rate is important for useful real-world applications.

Task 3: surveillance video summaries

As indicated in the anomaly detection test, most security camera footage is not of interest and reflects benign activities. An additional monitor need is therefore an automated video summarization capability which produces shortened but accurate summaries of long video files. The goal is a CV program that can reduce an eight-hour raw video, for example, to an edited 'change only' video lasting minutes which highlights interesting activities, eliminates redundant information, and improves camera usefulness for investigation purposes (Ashby 2017; Chen et al. 2009; Dowling et al. 2019; Evangelopoulos et al. 2009; Gao et al. 2009). The most common approaches for generating compact video summations comprise three phases. First, determination of the set of video regions which contain dynamic and moving objects is conducted. This is followed by detection of specific activities and actions of interest labeled by pre-trained classifiers (for example assaults would be noted and flagged). Finally, the video regions which contain events of interest are selected and the video summary is optimized to avoid redundancy and overlap, compressing the original video to its highlights. To this end, advanced techniques to index video features have been developed using compacting coding schemes (where just tens or hundreds of information bits per video feature are stored compared to millions of bits for non-summarized videos), that preserve visual information, and are useful for query-directed searches (Ye et al. 2013). The practical impact for a human monitor would be significant gains in search speed and the ability to search many thousands of hours of video data quickly for specific elements and events while providing significant reductions in storage and computational costs associated with multi-camera networks.

CV example: unsupervised video summations

A promising fully automatic video summation method was developed using semantic indexing for action and event detection and a neural network-based approach (see Ren et al. 2015) for the automatic detection of objects. For processing, a video was first summarized by generating distinctive video clips having the most discriminative information to create a unique visual dictionary. This dictionary was then used to score the non-dictionary clips in a video. A reconstruction score was used as a measure of distinctness, where higher scores indicate the importance (or difference) of each video segment from the balance of the origin video. The dictionary creation process was repeated until multiple video subsets that covered the entire length of the video were obtained and each video clip had been scored. Lastly, the video was summarized by selecting the highest scoring clips that reflected important events in the video and contained unique visual information. These summaries were then available for human operator review and queries.

An example of the unsupervised video summarization approach is shown in Fig. 6 for a video containing a car crash. For those frames containing the car

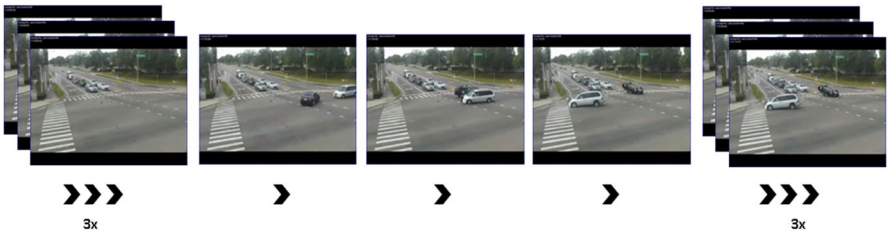


Fig. 6 Video time-lapse of unsupervised video summarization approach

accident, the original time-lapse and details of the crash were preserved. As the approach was tested on unannotated raw surveillance videos, there were no ground truth comparisons possible, nor were there available previous CV-based methods for empirical comparison. Therefore, only a qualitative assessment was available. For the fully automatic unsupervised summaries, the unsupervised video summation approach was deemed to perform well in the computer vision lab setting and correctly summarized raw videos into non-overlapping video segments. The approach not only functioned well unsupervised, but had the ability to incorporate pre-existing user defined events of interest, thereby offering the capability to camera operators to directly query security camera networks.

Task 4: query-directed searches of long video files

A final monitor task is the need to search large video files and retrieve segments with particular properties or visual traits such as a robbery or weapon. As an adjunct of video summations, in the tested query-directed search CV approach, the addition of user supplied information enabled the algorithm to better meet monitor needs. The CV software determination to add a video sequence to the final summary depended on both the sequence's relevance to the query and its uniqueness in the context of the entire video.

CV example: a CRCV query-focused video search

A monitor query-directed search method was developed and tested using two qualitatively different data sets.[4] Similar to high-quality surveillance videos more relevant to security operations, the first data set's videos were long and were recorded in an uncontrolled environment from a first-person view. As a result, many of the visual scenes are repetitive. In contrast, the second developmental data set was drawn from television episodes from a third person viewpoint and the scenes were controlled and concise. CV-generated video summaries were evaluated by contrasting them

[4] The first dataset, UT Egocentric (UTE), included four daily life egocentric videos, each 3–5 h long compiled by Ghosh, et al. (2012). The second data set of television episodes set contained four videos, each roughly 45 min long from Yeung et al. (2014).

Table 1 Comparison results for four query-generated video summarization methods

	UTE dataset (%)			TV episodes dataset (%)		
	F	Precision	Recall	*F*	Precision	Recall
Learning submodular mixtures (Gygli et al. 2015)	20.98	31.40	26.99	32.19	41.59	27.01
Quasi real-time summarization (Zhao and Xing 2014)	12.45	19.47	13.14	31.88	27.49	41.69
Determinantal point processes (Kulesza and Taskar 2012)	15.70	19.22	32.08	29.62	35.26	34.00
CRCV query directed	21.27	17.87	41.65	37.02	38.41	36.82

against a "ground truth" summary provided by human annotators with associated precision error, recall accuracy, and F-scores as reported in Table 1.[5] The shared goal for each approach was to maximize recall percentage and minimize precision error percentage. Table 1 shows the comparative results for three recent different CV video summarizers and the CRCV method. This approach outperformed the three alternate approaches, improved the rate at which unique events were discerned, and reduced the false-positive rate from the misclassification of events. Based on its higher average F-scores, the CRCV method generated better overall summaries. Importantly, CRCV was more accurate on the security-like video data.

Discussion

A set of CV programs developed with security applications in mind were tested in a CV lab. Relevant for both the monitoring of live video and the post hoc review of archived video files, CV algorithms for the detection of events of interest and anomalies and for querying and summarizing lengthy videos were developed and tested. CV approaches to these tasks showed promise in their lab assessments either outperforming alternative methods or approaching human-level ground truth performance levels. Table 2 summarizes the four monitoring tasks, tested CV approaches, and the implications of their development.

Looking beyond their lab assessments, Table 2 lists the possible impacts on security organizations from the development of successful CV applications. The primary expected benefits lie first in cost and personnel time savings for agencies. In addition to organizational benefits, an increase in the general effectiveness of camera networks will shift them to security tools that have better deterrent effects and produce more investigative leads and case evidence.

[5] Recall refers to the accuracy of a method in discerning the actual number of correct events in a video, and precision refers to the number of misclassified instances. For example, if a query was to identify police cars in a video that contained 10 police cars (its' ground truth), and 8 segments were identified as police cars, the recall percentage would be 8 of 10 or 80%. If 1 of 8 identified police cars were incorrect, the method's precision rate would be 1 of 8 or 12.5%. In practical terms, the goal is to maximize the recall rate and minimize the precision error.

Table 2 Four monitor tasks and computer vision

Security need	CV solution	Computer vision example	Lab assessment	Positive implications for security organizations	Pro and Con implications for society
Detection of crimes and suspects	Automatic action and event detection	Tube Convolutional Neural Network T-CNN	ROC curve improvement compared with two prior computer vision approaches	Increased proactive intervention, tracking, reduced monitor personnel needs, increased deterrence and clearance rates	Pro: decreased profiling and voyeurism Con: Net widening, decreased informal guardianship
Detecting and alerting to the unexpected	Anomaly detection	Weakly Labeled Supervised Learning	Ground truth comparison	Increased response to ongoing events, increased effectiveness of camera networks, more flexible automated surveillance, alerting to rare but important events	Pro: increased deterrent effects, increased real-time response Con: net widening
Storage and video file searches	Automatic generation of surveillance video summaries	Unsupervised Video Summation	Qualitative assessment	Cost, time and personnel savings, generation of investigation and evidence information, increased speed and accuracy in file searches, use in cold case investigations	Pro: increased effectiveness of formal guardianship Con: increased potential for creation of dossiers on low risk individuals
Finding specific scenes in long video files	Query-based searches	CRCV Query-Directed Summaries	Improvement in recall and precision over prior methods	Tracking of suspects across camera fields; cost, time, and personnel savings	Pro: increased clearance rates Con: Net widening

The limited research on surveillance cameras and investigations suggests that solutions for many crime types results from the availability of surveillance camera video (Ashby 2017). CV stands as a potentially significant investigation enhancer. As timely access to footage for criminal investigations has been associated with increased clearance rates (Morgan and Dowling 2019), an increase in speed of access to footage generated by CV enhancements should be associated with improved investigation outcomes and clearance rates. As CV increases the capturing of events of interest, a cycle of increased usefulness and subsequent increased investigator requests should ensue. CV also should increased investigative usefulness for crimes where video is reported as often not useful even when available such as cycle thefts, theft of vehicles, theft from vehicles, criminal damage, and theft from persons (Ashby 2017).

Camera footage also has been reported as particularly useful in the earlier stages of investigations, precisely where CV is most likely to help through noting previously missed events. And as camera footage is less likely to be available for crimes occurring at unknown times or remote locations (Ashby 2017), CV should assist in both situations by speeding file searches and automating the monitoring of low-activity areas. CV-enhanced networks would increase quick detection and automatically flag relevant video sequences and directly address the problem of long temporal windows for crimes resulting in fewer requests for footage noted by Ashby (2017).

To fully understand its' investigative potential, future research is needed on the interplay of CV, crime detection, timeliness, availability, and usefulness of surveillance cameras for investigations. Existing CV capabilities could increase requests for camera video and eliminate the need to wait on human investigator requests by automatically copying and sending video sequences of events of interest. In addition to investigations, active camera monitoring has been flagged as a significant factor in generating crime-reducing results (Piza et al. 2019). With CV, active monitoring does not require a human. By automating active monitoring, CV can increase and improve camera networks' utility for investigations. In sum, it is a reasonable expectation for CV to increase deterrence, detection, availability, and usefulness of surveillance camera networks. On the other hand, negative implication for organizations from successful but misused CV capabilities would be the waste of organizational resources and an increase in public distrust of monitoring. Before definitive conclusions can be forwarded, though, the CV/human investigator interface needs rigorous field research.

Related to the positive and negative implications for society, effective CV software can have positive impacts on a set of concerns that have been regularly raised regarding manually monitored surveillance camera networks (Welsh et al. 2015). Successful CV solutions can address concerns such as data swamping, boredom, profiling, and voyeurism and could make security camera networks true predictive tools. It would appear that CV-enhanced systems have the potential to increase the use of surveillance cameras for prevention and detection due to their increasing technological capabilities while simultaneously reducing the rate of missed criminal activity. The same technological enhancements, however, raise concerns that cannot be addressed by computer vision and may be exacerbated by it (Adams and

Ferryman 2015; Leman-Langlois 2002; Surette 2005). Net widening, suppression of citizen guardianship, oppressive use, expansion of data collection portfolios on low threat individuals, and increased social distance between the surveilled and the surveillers will remain concerns.

An associated issue is the loss of privacy (Adams and Ferryman 2015) and a concern of a pernicious psychological effect from a "voyeuristic gaze" on individuals and societies has long existed (see Marx 1988). As computer vision is likely to provide more vigilant and observant surveillance, freedom from surveillance is likely to be reduced. Beyond the effect from being better at identifying individuals, the issues of privacy and control of images are inherent in public space surveillance and can be significantly exacerbated by a CV-enhanced camera network. Discussing early human-monitored camera networks, Norris and Armstrong (1999) pointed out that "exclusionary surveillance" can occur from camera networks, an effect potentially enhanced if CV is used to identify and keep "types" of people out of specific public spaces. Associated with a loss of privacy is the concern that CV-enhanced camera systems will be used as tools of oppression (Graham 1996). CV will allow large camera networks to be more effectively monitored so that the number of subjects contained in a system's files and amount of data per subject will greatly increase. In gist, the development of CV-based capabilities likely will not relieve concerns about net widening, privacy, and oppression and will worsen them if misused (Adams and Ferryman 2015).

The degradation of informal citizen guardianship is also a general concern with CV cameras. The fear is that the technology will become a substitute for people and the natural surveillance that comes from human interaction (Surette 2006). The issue of reduction in informal citizen guardianship is not a trivial concern as the evaluations of non-CV-enhanced camera projects have generally concluded that crime reduction lies more in crime deterrence effects than in direct crime detection (Piza et al. 2014b; Welsh and Farrington 2009). The interplay between formal (police sourced) and informal (citizen sourced) guardianship is little understood (Surette 2006) and the role of camera networks and CV applications is unexplored. Whether computer-enhanced camera networks reduce citizen guardianship remains an important unexamined research question. Lastly, computer-enhanced surveillance will not reduce the likelihood of spatial displacement—offenders who can will likely move to non-surveilled areas. If CV camera networks are highly effective they may increase the concentration of crime and disorder in non-surveilled areas. However, diffusion of positive deterrent benefits from CV-enhanced cameras to nearby non-surveilled areas is also likely. Similar to effects on guardianship, the effect of CV-enhanced cameras on displacement of crime diffusion of benefits is an unexamined research question. Last, although the distancing between monitors and those monitored is not an inherent effect of CV, depending upon how the technology is utilized such an effect is a possibility. If the enhanced systems are employed so as to increase "surveillance at a distance", relationships between agencies and the public in marginalized communities would be further strained.

In sum, on the positive side, appropriately used CV camera networks would decrease profiling, formal guardianship would be strengthened, real-time response to incidents will be increased, and usefulness for investigations will be enhanced.

The primary possible negative effect from successful CV lies in net widening where minor offenders would be more often swept up into the criminal justice system. The degradation of informal citizen guardianship and public trust are secondary potential negative effects. None of these possible benefits or concerns could be assessed in the laboratory.

What do these findings imply for the future use of the millions of existing security cameras? Prior research has found a general positive view toward surveillance cameras in the form of neighborhood, vehicle mounted, and body-worn cameras (Sandhu 2017), and an initial assessment of police officers reported similar results regarding CV-enhanced cameras (Shah 2017). The key change that CV will prompt will be the shifting of humans from camera monitors to camera network supervisors. The former must watch multiple camera feeds or scan video files for long hours, the latter will not have to watch any camera feeds and need only be available to review CV-flagged video sequences and render response decisions. In a CV-enhanced world, a human network supervisor could reasonably and accurately oversee hundreds of cameras as compared to the long-standing functional limit of 5–15 cameras that can be reliably humanly monitored (Tickner and Poulton 1973). CV could revolutionize how surveillance cameras networks are manned and supervised and if successfully utilized, CV can turn camera networks from haphazard, post hoc, hit-or-miss security tools into effective real-time effective detection and deterrent systems.

Limitations

Limitations of this research are twofold. First and primary, the results reported herein are all based on computer vision laboratory tests. The history of security and technology is filled with instances of promising technology failing when transferred from controlled laboratory settings to uncontrolled field environments. The fact that the CV algorithms performed better than prior approaches or approached human-level accuracy levels suggests but does not guarantee that they will perform adequately when installed in real-world camera monitoring rooms. The interface between humans, the CV technology, and monitoring tasks has not been addressed. A second limitation was the limited number of training video examples that were available for computer vision program development. Some events of interest such as robberies are rarely captured on surveillance cameras and it proved difficult to locate sufficient numbers of examples to train CV program classifiers. In addition, events sometimes occurred in conjunction with other activities or were ambiguous. Therefore, the difficulty in training event detectors varied significantly across the types of events that were of interest.

Conclusion

Computer vision can shift the current reality of camera surveillance toward what the public wants, increased safety, better-quality investigations, and less disorder, and away from what the public has often received, increased but haphazard monitoring

and more arrests for less serious offenses. The ultimate technological goal is to have cameras with embedded CV capabilities to create intelligent decentralized camera systems where, in addition to being part of a large network, each security camera will have an independent self-analysis capability. The impact of such capabilities on agencies and society is both promising and concerning.

If CV results in event swamping from the flagging of numerous events for review or conversely has no significant impact on daily security operations, the promise of computer vision will be unmet. To determine whether its promise is delivered, the transfer from laboratory to field is necessary. Whether CV-enhanced cameras will result in more effective use of security camera networks that have been heavily invested in is unknown, but the technological promise exists to explore.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Abdi, H., D. Valentin, and B. Edelman. 1999. *Neural networks*. Thousand Oaks, CA: Sage.

Adams, A., and J. Ferryman. 2015. The future of video analytics for surveillance and its ethical implications. *Security Journal* 28 (3): 272–289.

Alexandrie, G. 2017. Surveillance cameras and crime: A review of randomized and natural experiments. *Journal of Scandinavian Studies of Criminology and Crime Prevention* 18 (2): 210–222.

Andrews, S., I. Tsochantaridis, and T. Hofmann. 2003. Support vector machines for multiple-instance learning. In *Advances in neural information processing Systems*, 577–584. Cambridge: MIT.

Ashby, M.P. 2017. The value of CCTV surveillance cameras as an investigative tool: An empirical analysis. *European Journal on Criminal Policy and Research* 23 (3): 441–459.

Baldwin, D.A., and J.A. Baird. 2001. Discerning intentions in dynamic human action. *Trends in Cognitive Sciences* 5: 171–178.

Barrett, H., P. Todd, G. Miller, and P.W. Blythe. 2005. Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior* 26: 313–331.

Bredemeier, K., and D. Simons. 2012. Working memory and inattentional blindness. *Psychological Bulletin Review* 19: 239–244.

Bulwa, D., and M.B. Stannard. 2007. Is it worth the cost? *San Francisco Chronicle*, August 17. https://www.sfgate.com/news/article/Is-it-worth-the-cost-2546948.php. Downloaded 8 Oct 2019

Chen, B.W., J.-C. Wang, and J.F. Wang. 2009. A novel video summarization based on mining the story-structure & semantic relations among concept entities. *IEEE Transactions on Multimedia* 11 (2): 295–312.

Coetzer, B., B. Josephs, and J. van der Merwe. 2011. *Information management and video analytics: The future of intelligent video surveillance*. Rijeka: INTECH Open Access Publisher.

Davenport, J. 2007. Tens of thousands of CCTV cameras, yet 80% of crime unsolved. *Evening Standard*, September 19. https://www.standard.co.uk/news/tens-of-thousands-of-cctv-cameras-yet-80-of-crime-unsolved-6684359.html. Downloaded 8 Oct 2019.

Donald, C. 2005. How many monitors should a CCTV operator view. *CCTV Image*, Spring, 35–36.

Dowling, C., A. Morgan, A. Gannoni, and P. Jorna. 2019. How do police use CCTV footage in criminal investigations? *Trends and Issues in Crime and Criminal Justice* 575: 1–14.

Edwards, R. 2008. Police say CCTV is an 'utter fiasco'. *The Telegraph*, May 6.https://www.telegraph.co.uk/news/uknews/1932769/Police-say-CCTV-is-utter-fiasco-as-most-footage-is-unusable.html. *Downloaded 8 Oct 2019.*

Edwards, R. 2009. Seven of ten murders solved by CCTV. *The Telegraph*, January 1.https://www.telegraph.co.uk/news/uknews/law-and-order/4060443/Seven-of-ten-murders-solved-by-CCTV.html. *Downloaded 8 Oct 2019.*

Evangelopoulos, G., A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis, 2009. Video event detection and summarization using audio, visual and text saliency. In *ICASSP IEEE international conference on acoustics, speech and signal processing*.

Faber, L., N. Maurits, and M. Lorist. 2012. Mental fatigue affects visual selective attention. *PLoS ONE* 710: e48073.

Ferguson, A. 2017. Policing predictive policing. *Washington University Law Review* 94: 1115–1194.

Fougnie, D., and R. Marois. 2007. Executive working memory load induces inattentional blindness. *Psychonomic Bulletin and Review* 141: 142–147.

Gao, Y., D. Wang, J. Yong, and H. Gu. 2009. Dynamic video summarization using two level redundancy detection. *Multimedia Tools and Applications* 422: 233–250.

Gerell, M. 2016. Hot spot policing with actively monitored CCTV Cameras. *International Criminal Justice Review* 24 (2): 187–201.

Ghosh, J., Y.J. Lee, and K. Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *IEEE conference on CV and pattern recognition*, Providence, RI, pp. 1346–1353.

Gill, M. 2003. *CCTV*. Leicester: Perpetuity Press.

Gong, S., C.C. Loy, and T. Xiang. 2011. Security and surveillance. In *Visual analysis of humans*, 455–472. London: Springer.

Goold, B. 2004. *CCTV and policing*. Oxford: Oxford University Press.

Gowsikhaa, D., S. Abirami, and R. Baskaran. 2014. Automated human behavior analysis from surveillance videos: A survey. *Artificial Intelligence Review* 42 (4): 1–19.

Graham, S. 1996. CCTV-Big Brother or friendly eye in the sky? *T AND CP* 65: 57–59.

Gygli, M., H. Grabner, and L. Van Gool. 2015. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Hesse, L. 2002. The transition from video motion detection to intelligent scene discrimination and target tracking in automated video surveillance systems. *Security Journal* 15 (2): 69–78.

Hier, S., J. Greenberg, K. Walby, and D. Lett. 2007. Media, communication and the establishment of public camera surveillance programmes in Canada. *Media, Culture, and Society* 295: 727–751.

Honovich, J. 2008. Is public CCTV effective? July 7. https://ipvm.com/reports/is-public-cctv-effective. Downloaded 27 Sept 2019.

Hyman, I., E. Boss, S. Matthew, B. Wise, M. McKenzie, E. Kira, and J. Caggiano. 2009. Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone. *Applied Cognitive Psychology* 245: 597–607.

Idrees, H., Shah, M., & Surette, R. 2018. Enhancing camera surveillance using computer vision: A research note. *Policing: An International Journal* 41 (2), 292–307.

Jain, M., J. Van Gemert, H. Jégou, P. Bouthemy, and C.G. Snoek. 2014. Action localization with tubelets from motion. In *Proceedings of the IEEE conference on CV and pattern recognition*, pp. 740–747.

Keval, H., and M. Sasse. 2010. "Not the usual suspects": A study of factors reducing the effectiveness of CCTV. *Security Journal* 232: 134–154.

Kuehne, H., H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: a large video database for human motion recognition. In *International conference on CV*, pp. 2556–2563.

Kulesza, A., and B. Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning* 5 (2–3): 123–286.

La Vigne, N., S. Lowry, J. Markman, and A. Dwyer. 2011. *Evaluating the use of public surveillance cameras for crime control and prevention*. Washington, D.C.: Urban Institute, Justice Policy Center. https://www.urban.org/UploadedPDF/412403-Evaluating-the-Use-of-Public-Surveillance-Cameras-for-Crime-Control-and-Prevention.pdf.

Leman-Langlois, S. 2002. The myopic panopticon: The social consequences of policing through the lens. *Policing and Society* 131: 43–58.

Mack, A., and I. Rock. 1998. *Inattentional blindness*. Cambridge, MA: MIT Press.

Marx, G. 1988. *Undercover: Police surveillance in America*. Berkley: University of California Press.

Memmert, D. 2006. The effects of eye movement, age, and expertise on inattentional blindness. *Consciousness and Cognition* 153: 620–627. https://doi.org/10.1016/j.concog.2006.01.001.

Morgan, A., and M. Coughlan. 2018. Police use of CCTV on the rail network. *Trends and Issues in Crime and Criminal Justice* 561: 1–17.

Morgan, A., and C. Dowling. 2019. Does CCTV help police solve crime? *Trends and Issues in Crime and Criminal Justice* 576: 1–14.
Most, S.B., B.J. Scholl, E.R. Clifford, and D.J. Simons. 2005. What you see is what you set: Sustained inattentional blindness and the capture of awareness. *Psychological Review* 1121: 217–242.
Näsholm, E., S. Rohlfing, and J.D. Sauer. 2014. Pirate stealth or inattentional blindness? The effects of target relevance and sustained attention on security monitoring for experienced and naïve operators. *PLoS ONE* 9 (1): e86157. https://doi.org/10.1371/journal.pone.0086157.
Norris, C., and G. Armstrong. 1999. *The Maximum Surveillance Society: The rise of CCTV*. Oxford: Berg.
Piza, E., J. Caplan, and L. Kennedy. 2014a. CCTV as a tool for early police intervention: Preliminary lessons from nine case studies. *Security Journal* 30: 247–265. https://doi.org/10.1057/sj.2014.17.
Piza, E., J. Caplan, and L. Kennedy. 2014b. Is the punishment more certain? An analysis of CCTV detections and enforcement. *Justice Quarterly* 31 (6): 1015–1043.
Piza, E., B. Welsh, D. Farrington, and A. Thomas. 2019. CCTV surveillance for crime prevention: A 40-year systematic review with meta-analysis. *Criminology & Public Policy* 18: 135–159.
Prenzler, T., and E. Wilson. 2019. The Ipswich (Queensland) safe city program: an evaluation. *Security Journal* 32: 137–152.
Ratcliffe, J.H., T. Taniguchi, and R.B. Taylor. 2009. The crime reduction effects of public CCTV cameras: a multi-method spatial approach. *Justice Quarterly* 264: 746–770.
Ren, S., K. He, R. Girshick, and J. Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99.
Sandhu, A. 2017. 'I'm glad that was on camera': A case study of police officers' perceptions of cameras. *Policing and Society* 29 (2): 223–235.
Sasse, A. 2010. Not seeing the crime for the cameras? *Communications of the ACM* 53: 22–25.
Scheitle, C.P., and C. Halligan. 2018. Explaining the adoption of security measures by places of worship: Perceived risk of victimization and organizational structure. *Security Journal* 31 (10): 1–23.
Shah, M. 2017. Project Report: Studying the impact of video analytics for pre, live and post event analysis on outcomes of criminal justice, July 2016–December 2016. Orlando, FL: University of Central Florida Center for Research on Computer Vision. Funded by U.S. Department of Justice, NIJ-2015-R2-CX-K025.
Surette, R. 2005. The thinking eye: Pros and cons of second generation CCTV surveillance systems. *Policing: An International Journal of Police Strategies and Management* 281: 152–173.
Surette, R. 2006. CCTV and citizen guardianship suppression: A questionable proposition. *Police Quarterly* 9: 100–125.
Surette, R. 2015. *Media, crime, and criminal justice: Images, realities, and policies*. Stamford, CT: Cengage.
Taylor, E. 2010. Evaluating CCTV: Why the findings are inconsistent, inconclusive and ultimately irrelevant. *Crime Prevention and Community Safety* 124: 209–232.
The Scotsman. 2008. CCTV: Does it actually work? The Scotsman, May 28. https://www.scotsman.com/news-2-15012/cctv-does-it-actually-work-1-1169849. Downloaded 8 Oct 2019.
Thomas, J., and K. Cook. 2006. A visual analytics agenda. *IEEE Computer Graphics and Applications* 261: 10–13.
Tickner, A., and E. Poulton. 1973. Monitoring up to 16 synthetic television picture showing a great deal of movement. *Ergonomics* 16: 381–401.
Uijlings, J.R., K.E. Van De Sande, T. Gevers, and A. Smeulders. 2013. Selective search for object recognition. *International Journal of CV* 1042: 154–171.
Wang, L., Y. Qiao, and X. Tang. 2014. Video action detection with relational dynamic-poselets. In *European conference on CV*, pp. 565–580. Cham: Springer.
Welsh, B., and D. Farrington. 2002. *Crime prevention effects of closed circuit television: A systematic review*. Home Office Research Study 252. London: Home Office.
Welsh, B., and D. Farrington. 2004. Evidence-based crime prevention: The effectiveness of CCTV. *Crime Prevention and Community Safety* 6: 21–33.
Welsh, B., and D. Farrington. 2009. Public area CCTV and crime prevention: An updated systematic review and meta-analysis. *Justice Quarterly* 264: 716–745.
Welsh, B., D. Farrington, and S. Taheri. 2015. Effectiveness and social costs of public area surveillance for crime prevention. *Annual Review of Law and Social Science* 11: 111–130.
Yang, M., S. Ji, W. Xu, J. Wang, F. Lv, K. Yu, Y. Gong, M. Dikmen, D.J. Li, and T.S. Huang. 2009. Detecting human actions in surveillance video. In *TREC video retrieval evaluation workshop*.

Ye, G., D. Liu, J. Wang, and S. Chang. 2013. Large-scale video hashing via structure learning. In *Proceedings of the IEEE international conference on CV*, pp. 2272–2279.

Yeung, S., A. Fathi, and L. Fei-Fei. 2014. Videoset: Video summary evaluation through text. arXiv preprint. arXiv:1406.5824.

Zhao, B., and E.P. Xing. 2014. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Zhou, Z.H. 2018. A brief introduction to weakly supervised learning. *National Science Review* 5 (1): 44–53.