PIJPSM
41,2

Received 3 November 2016
Revised 13 March 2017
29 March 2017
Accepted 4 April 2017

Enhancing camera surveillance using computer vision: a research note

Haroon Idrees and Mubarak Shah
Center for Research in Computer Vision, University of Central Florida, Orlando, Florida, USA, and
Ray Surette
Department of Criminal Justice, University of Central Florida, Orlando, Florida, USA

Abstract

Purpose – The growth of police operated surveillance cameras has out-paced the ability of humans to monitor them effectively. Computer vision is a possible solution. An ongoing research project on the application of computer vision within a municipal police department is described. The paper aims to discuss these issues.

Design/methodology/approach – Following the demystification of computer vision technology, its potential for police agencies is developed within a focus on computer vision as a solution for two common surveillance camera tasks (live monitoring of multiple surveillance cameras and summarizing archived video files). Three unaddressed research questions (can specialized computer vision applications for law enforcement be developed at this time, how will computer vision be utilized within existing public safety camera monitoring rooms, and what are the system-wide impacts of a computer vision capability on local criminal justice systems) are considered.

Findings – Despite computer vision becoming accessible to law enforcement agencies the impact of computer vision has not been discussed or adequately researched. There is little knowledge of computer vision or its potential in the field.

Originality/value – This paper introduces and discusses computer vision from a law enforcement perspective and will be valuable to police personnel tasked with monitoring large camera networks and considering computer vision as a system upgrade.

Keywords Crime prevention, Computer vision, Community policing, Police cameras, Public space surveillance cameras

Paper type Conceptual paper

Introduction

Police surveillance of public spaces

Historically called the "stake-out," police surveillance has a long history and evidence gained from surveillance has been an important part of investigations for nearly two centuries (Marx, 1988). Similarly, the use of visual technology by police began in the nineteenth century with the photographing of inmates and evolved to include crime scene photographs as standard police procedures (Buckland, 2001; Norris and Armstrong, 1999a). While both practices became law enforcement mainstays, police surveillance and visual evidence remained separate realms well into the twentieth century (Norris and Armstrong, 1999a). Surveillance cameras operated by law enforcement are therefore a relatively recent phenomenon and their marriage has moved surveillance from a human based activity to a heavily technological one.

As evolved, police surveillance has two goals: proactively deterring offenders and aiding in investigations. Initially, the investigative goal dominated and surveillance was aimed at solving crimes, not preventing them. But as cameras became less expensive and more

Funded by US Department of Justice National Institute of Justice, NIJ-2015-R2-CX-K025.

Policing: An International Journal of Police Strategies & Management
Vol. 41 No. 2, 2018
pp. 292-307

1363-951X
DOI 10.1108/PIJPSM-11-2016-0158

pervasive, deterrence and risk reduction became important (Kroener, 2014). Current camera surveillance projects aim to provide some combination of retrospective crime scene analysis, deterrence of future crimes, and facilitation of real-time intervention and force deployment (Haggerty and Gozso, 2005).

It is unknown how many public space surveillance cameras are operated by law enforcement agencies but a 2014 US estimate was about 30 million (Staples, 2014, p. 71 citing Vlahos, 2009). Despite their limitations, surveillance cameras have emerged as a popular law enforcement choice to address crime and security concerns and much of the gap between what was promised and what was delivered has been linked to their rapid adoption (Surette, 2005)[1]. The number of cameras installed quickly outpaced the capacity to monitor them and thus to effectively respond to what was visually captured (Piza *et al.*, 2014a, b; Gill *et al.*, 2005; Keval and Sasse, 2010). It is apparent that a weak link in the information chain from camera to police response is the human monitor tasked with watching the screens (Surette, 2005).

Humans as camera monitors: a poor match

Humans are not particularly good as camera monitors (Hier *et al.*, 2007; Näsholm *et al.*, 2014; Sutton and Wilson, 2004). Surveillance camera monitors are most frequently tasked with general camera monitoring. They sit at a desk before a bank of monitor screens and conduct on-going non-specific assessment of live video feeds. The review of a surveillance camera's archived video is also sometimes required to determine if an event of interest was recorded. In this second task, monitors are asked to search for a specific event, person, or object. Again, the human monitor is often asked to watch hours of video. The monitoring difficulty is further compounded in that many criminal activities have subtle precursors that are easily overlooked when humans are tasked with monitoring multiple cameras (La Vigne *et al.*, 2011; Piza *et al.*, 2014a; Piza and Sytsma, 2016). Humans quickly become image swamped, missing more than they observe even when vigilant (Boksem *et al.*, 2005; Faber *et al.*, 2012; Gill *et al.*, 2005; Sasse, 2010; Surette, 2005).

The deficiency of human monitors occurs because perception failure occurs when there is little visual change present in long video stretches. The monitor's attention shifts from visual review to other non-visual tasks such as conversing or daydreaming resulting in "inattentional blindness" (Johnston *et al.*, 1990; Sasse, 2010). In these instances, monitors have their eyes open and are looking at a video stream but their minds are cognitively elsewhere, the visual images failing to reach psychological "attention capture" levels necessary for effective monitoring[2]. Significant amounts of time and video can pass, the images passing in plain view but unseen (Driver, 1998).

Relevant for specific event searches, perceptual blindness more often occurs when a monitor's cognitive attention is focused on finding one type of activity to the exclusion of other significant events (Bredemeier and Simons, 2012; Fougnie and Marois, 2007; Most *et al.*, 2005). In this situation, unexpected and even bizarre events are more likely to fail to capture the attention of monitors. Important for noting anomalies, such as unexpected crimes, when a monitor is looking intently for a particular element in a video stream, failure to see other things of interest increases (see, e.g. Piza *et al.*, 2014a, p. 10). The more different the unexpected event is from what is being looked for, the more likely it is to be missed (Memmert, 2006)[3]. Hence, different but serious crimes than one being searched for are less likely to be noticed, the opposite of the case in general monitoring tasks where the lack of visual change contributes to monitor error. In addition to these cognitive barriers, a number of surveillance barriers that reduce potential deterrent effects from surveillance cameras have been described. In addition to defensive actions taken by offenders (Piza and Sytsma, 2016) and camera-related contextual factors (Lim and Wilcox, 2017), two additional noted barriers are high camera to operator ratios (Piza *et al.*, 2014a, b, 2015) and poor integration into agency practices (La Vigne *et al.*, 2011; Piza *et al.*, 2014b, 2015).

The cumulative result is that "a high camera-to-operator ratio has the predictable result of crime occurring within sight of a camera going undetected and the detection of criminal events by CCTV operators as rare" (Piza *et al.*, 2014a, pp. 1019-1020 citing Norris and Armstrong, 1999a, b). Faced with significant competition for attention, the camera systems currently are "hit or miss" tools regarding the detection of on-going incidents and expensive time and human capital consuming drudgery-laden search platforms for finding useful investigative evidence. Additionally criticized for using sworn personnel as monitors and for instances of monitor abuses such as voyeurism and profiling, for police agencies the need for an alternative to human monitors is apparent (Bredemeier and Simons, 2012; Surette, 2005). Computer vision applications can potentially address these issues and increase the deterrent impact of cameras and their organizational benefits. However, the lack of computer vision use in law enforcement is exacerbated by a lack of computer vision software development designed with law enforcement needs in mind and the absence of field trials to justify agency costs for upgrading to computer vision enhanced camera networks.

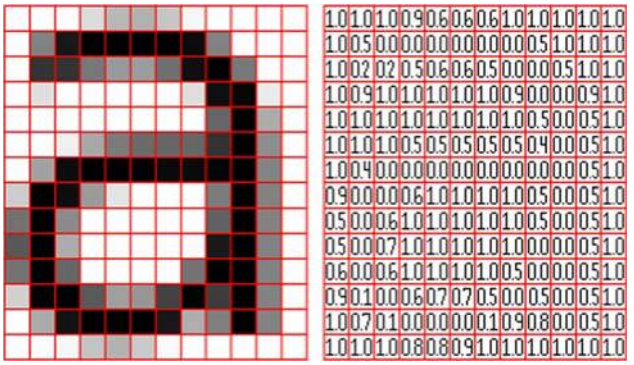
Computer vision as solution

An emerging approach to the shortfalls of human monitored camera systems is computer vision (also known as machine vision). A literature search reveals little current use of computer vision capabilities by law enforcement agencies although calls for its incorporation and discussions of potential applications have been forwarded (see Baldwin and Baird, 2001; Barrett *et al.*, 2005; Piza *et al.*, 2014b; Shah *et al.*, 2007; Thomas and Cook, 2006). The bulk of law enforcement computer applications have concentrated not on computer vision but on data-mining coupled with crime mapping to identify crime "hot spots" (Lohr, 2012; Wang *et al.*, 2013; Yu *et al.*, 2011). The common current computer vision uses are facial recognition applications and license plate readers. While computer vision as a public safety tool remains under-explored, the recent coupling of surveillance cameras to fast, inexpensive computers have made computer vision solutions feasible. The primary benefit that computer vision offers law enforcement agencies is the substitution of automated analysis of camera video streams for human monitors. With computer vision, the human in a computer vision enhanced security camera network assumes a supervisory assessment and response decision role.

The first step in understanding computer vision involves comprehending digitization of a visual image into a grid of pixels where each pixel is assigned a numerical value representing its color. This initial process generates for an image (or in the case of a video, each frame) a two-dimensional grid of numbers which mathematically renders the original image as digits, hence the term "digital photo." A simplified example is provided in Figure 1. These assigned numbers are the foundation for all subsequent manipulation, analysis, interpretation, labeling and other higher-level computer vision capabilities. When a digital photo is opened for viewing, the process is reversed by a photo processing program which uses each pixel's value to instruct the viewing device (a computer, smart phone, or digital camera) on how to color a corresponding screen pixel – converting numbers back to colored pixels and reconstructing the picture in a form that humans can see. This "picture to number to picture" process makes computer vision possible.

The key to computer vision is the analysis made possible by the pixel values when the state of the image is not visual but numerical. The art of computer vision moves quickly from input that looks loosely analogous to an "image" (e.g. the numbers assigned to each pixel in Figure 1) to working with data and outputs that do not appear to correspond to the original image in any straightforward fashion as the photo and histogram in Figure 2 demonstrate.

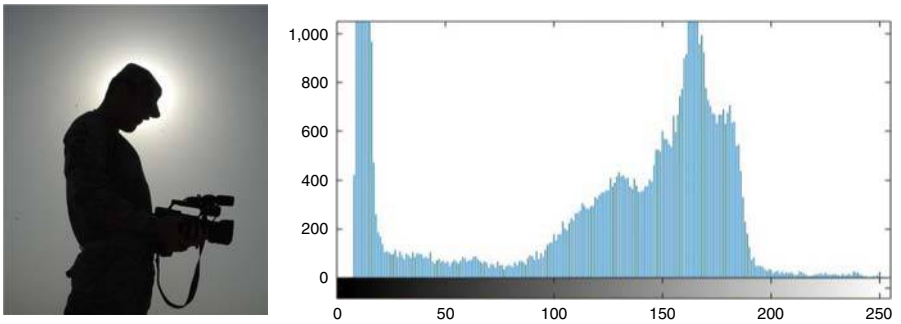
In essence, computer vision involves determining what a quantitative analysis of pixel values can tell about an image. From its numerical foundation the core tasks of computer vision proceed and in turn allow the development of common computer vision applications such as locating people and places in images, face recognition, and image stabilization.



Notes: The original letter "a" is rendered into a 14 by 12 array of 168 pixels with each pixel assigned a value representing a tone from "white" scored 1.0 to black squares scored 0.0. Gray squares are scored from light to dark (gray squares scored 0.5, light gray ones scored from 0.01 to 0.04, dark gray squares scored from 0.6 to 0.9). The digitized photo of the original "a" image results where the "a" can be vaguely discerned in the pattern of 0.0 scored pixels. For capturing color, multiple values, typically corresponding to red, green and blue, are stored for each pixel location

Source: Available at: http://pippin.gimp.org/image_processing/images/sample_grid_a_square.png

Figure 1.
A digitized letter



Notes: The histogram on the right is constructed from grayscale pixels from the image by quantizing them into 256 bins. The significance of image histograms for computer vision is that histograms are often unique for different objects and correlate with their shape, texture and color and can be used for assigning labels to images

Source: Photograph by SSG Brien Vorhees (public domain), via Wikimedia commons

Figure 2.
An image histogram

An example of a core computer vision task with direct criminal justice applications is the tracking of objects across a set of video frames (Yilmaz *et al.*, 2006). The mathematical representations of the tracked objects are derived from determining key points (unique sets of pixels in an image) so that a "probability distribution function" (PDF) can be generated. A PDF is analogous to a visual fingerprint without being as individually unique. Thus, an object will usually have a similar PDF that can be tracked across video frames. The objects

tracked can be automatically set by a computer vision object detector or can be assigned manually by a human placing an outline around a region of interest within a video frame. Recent tracking methods can lose and reacquire objects as they move into and out of camera fields of view (Comaniciu *et al.*, 2000) and work well when the tracked object is substantially different from its background. Multiple objects that are similar in appearance or that cross in front or behind one another are more difficult to track[4].

Another useful computer vision task is the assignment of a name to objects and actions. Not only is it useful to be able to name objects in a picture but it is an important goal to determine whether a human is present and to determine who they are and what they are doing, a crucial public safety surveillance task. For a computer vision program to be able to recognize and name objects, a set of images are initially used to train "classifiers," computer vision sub-routines that assign labels to images. Once developed, classifiers can be used to answer queries about unlabeled images such as: Is there a handgun in this video? The classifier training process for a previously unknown object proceeds after manual annotation during which training data are created by humans who assign labels to a set of representative images of the object (positive visual examples as shown in Plate 1) as well as those which do not contain the object (negative examples). The training images allow the computer vision program to mathematically differentiate among objects. For example, providing examples of "weapons" and "not weapons" trains a classifier that can then better calculate the probability of a weapon being in an image. Such an approach is termed "supervised" as it assumes availability of annotated training data in contrast to "semi-supervised" and "unsupervised" approaches that require partially annotated or no training data, respectively. In general, the performance of a particular computer vision task is proportional to the amount of human-labeled data available.

The now common task of matching a particular individual's face with a face in an image database is also useful (Turk and Pentland, 1991). To accurately match a face, the computer vision program must consider "between class variation" (different people who share features such as blue eyes) and "within class variation" (the same person who looks differently due to differences in image aspects (frontal face compared with profile for example). As with object and action labeling, face recognition is set as a probabilistic outcome with some threshold level of image similarity needed to be reached before a match is declared. A recent improvement works to maximize the distinction between the faces of different persons and takes into account differences (i.e. normalizes within class variation) observed across multiple images of the same person[5]. With these and other capabilities under development, potential computer vision solutions for police surveillance camera tasks are now within reach.

Computer vision applications in policing

General law enforcement surveillance needs that computer vision can address fall under two umbrellas. The first set revolves around the need for automated real-time, live video stream analysis. The second involves the need for *post-hoc* searches of archived video files. Computer vision based automated identification of public safety events of interest addresses the first task and query-based searches of video files addresses the second.

Related to the first task, live real-time video analysis involves the need to rapidly identify and correctly respond to ongoing incidents. This capability is important because effectiveness of a surveillance system in reducing crime has been linked to real-time intervention. Unless surveillance results in someone showing up to address an observed problem, camera deterrent effects wane (Ariel, 2016; Gill, 2003; Goold, 2004; Piza *et al.*, 2014a, b; Welsh and Farrington, 2002). The development of live event analysis is conducted along two computer vision paths: action and event detection of pre-identified events of interest and detection of anomalous new, unanticipated but potentially noteworthy events.

Sources: Burglary, available at: www.youtube.com/watch?v=ZtqHcNAc7kE, Salt Lake City Police Department, Creative Commons CC BY; Assault, available at: www.youtube.com/watch?v=k8CPg6mFFpo, South Australia Police, Creative Commons CC BY; Vandalism, available at: www.youtube.com/watch?v=_DXIAKJpv04, Creative Commons CC BY

Plate 1.
Positive training examples for assault, theft, vandalism, and robbery

For real-time detection of activities, it is imperative that the system analyzes the surveillance video as it is captured and classifies actions and events as they appear. Of particular interest to public safety monitors are many activities which occur infrequently and are precursors to criminal activity (e.g. "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 because first they are rare and therefore have a limited number of examples available for analysis and second, they can be ambiguous and difficult to define mathematically. Hence, a murderous assault will likely occur only once in the lifetime of a camera's view-shed but it is crucial that it be noted by a computer vision program and that it be distinguished from one person giving another a vigorous friendly hug. Humans quickly distinguish the two activities; however, computer vision programs must be quantitatively trained to do so. To be useful, anomaly models also must continuously update and incorporate environmental changes, for instance changes in weather, crowd density, or lightning conditions at different times of the day.

Computer vision can also reduce the immense amount of time currently spent reviewing and searching videos. Even when it is known that a video contains specific images such as weapons, the minutes or seconds of interest are often buried within hours of output. A computer vision solution to this issue is query-based searches. To be useful, query-based searches require search options that permits retrieval of objects with particular properties such as a person with specific height, weight, race, gender, or appearance; or "objects" such as an item a person was carrying like an umbrella or back-pack. The ability to submit an object and attribute-based search would significantly reduce the number of irrelevant video clips that an investigator must review. Independent of specific query-linked searches, it is also useful to have computer vision based video summarization programs for the distillation of videos into shortened but accurate summaries. An eight-hour video can typically be reduced to an edited "change only" video lasting minutes (Chen *et al.*, 2009; Evangelopoulos *et al.*, 2009; Gao *et al.*, 2009).

In another potential use of computer vision, recent criminal justice research has used camera footage to study pre-crime visual cues. For example, Piza *et al.* (2014b) and Levine *et al.* (2011) used video footage to examine violence precursors and Moeller (2016) and Piza and Sytsma (2016) searched for correlates of illegal drug sales. Computer vision has the potential to significantly aid these research efforts and increase the use of surveillance videos as a data source. A number of prior research efforts have employed surveillance video as data (see Piza and Sytsma, 2016; Piza *et al.*, 2014a; Sampson and Raudenbush, 1999; St Jean, 2007) but their usefulness has been limited by heavy processing and time demands. Despite having a number of years of video, Piza and Sytsma (2016) had to limit analysis to a single year and 62 incidents due to processing workload; in their study, each minute of video equaled 20 minutes of transcription time. Lastly, as implied by Piza and Sytsma (2016) and Moeller (2016) a set of criminological theories and concepts such as routine activities, environmental crime, crime displacement, and hotspot analysis could benefit from the exploration of computer vision generated data.

On-going research

A National Institute of Justice funded study is underway to address three research questions associated with police use of computer vision (Shah *et al.*, 2015). In this study computer vision analytics for a large surveillance camera network is being developed and their integration into a Public Safety Visual Analytics Workstation (PSVAW) within a municipal police department will be field tested. The law enforcement targeted computer vision analytics under development include the retrieval of objects, concepts and events (Mazaheri *et al.*, 2015); the localization of actions in long untrimmed videos (Soomro *et al.*, 2016); the interactive detection of anomalies without annotated training examples (Zavesky and Chang, 2008); and multiple methods for video summarization (Rodriguez, 2010). The research questions addressed are:

RQ1. What is the accuracy and speed of the analytics?

RQ2. What is the organizational fit of computer vision in a police department?

RQ3. What is the impact of a computer vision capability on a municipal criminal justice system?

Research question 1. How well do the computer vision algorithms work in the lab? Computer vision algorithms are being evaluated on standard pre-curated, annotated data sets which are partitioned for training and testing. For many computer vision tasks, prior algorithm accuracy has been high, above 90 percent for easy action recognition data sets. However, for challenging data sets accuracy drops to around 60 percent (Kuehne *et al.*, 2011), a level that would generate numerous false hits in police applications. The goal is to produce computer vision algorithms that are sensitive enough to not miss significant events but also do not swamp human reviewers with large number of erroneously flagged video clips. A second programming goal is to achieve significant reduction in storage and computational cost over large-scale surveillance video archives. The practical impact for a law enforcement agency would be significant gains in search speed and the ability to search thousands of hours of video data (Ye *et al.*, 2013).

A computer vision based method for detection of static concepts and dynamic events is also being developed for object detection such as weapons, police officers, police vehicles; and complex event detection like assaults, thefts, and car crashes. Both use features from deep neural networks for processing images and video frames[6]. In the static concept search, a human can query a single concept such as "police officer" and the system will return a sorted list of video clips in which the concept "police officer" appears. The complex event detection categorizes video into broad categorical classes of behaviors beyond a brief appearance of single objects. Thus, more challenging activities can be dealt with and video clips can be robustly classified into events such as "robberies" and "assaults."

Regarding the need for real-time video analysis, computer vision software for live online abnormality detection is additionally being created (Roshtkhari and Levine, 2013). The quantitative problem amounts to finding patterns in the digital data that significantly deviate from behaviors previously identified empirically. The detection of abnormal behaviors is a difficult task. First, the quantitative definition of a normal vs abnormal visual pattern is not well defined. Second, normal behavior evolves over time and may change significantly as time passes (for instance, many people walking during daylight vs few people walking during nighttime differ visually but both may be normal activity when it comes to crime detection Lim and Wilcox, 2017; Moeller, 2016). Third, because abnormal events are rare it is difficult to obtain enough examples to train classifiers.

To cope with these challenges, an online dictionary learning approach to detect abnormalities is being pursued which divides long videos into small non-overlapping meaningful clips. Since these segments are computed based on appearance and motion information, many will contain tracked vehicles and people, which are then compared with existing elements in a dictionary thereby permitting the detection of abnormalities (Tran *et al.*, 2015). If the flagged anomaly is deemed a normal event, it is added to the dictionary. This allows the computer program to interactively update and recognize a "new normal" such as when a crowded day time street becomes a sparsely populated nighttime scene. The anomaly detection process flags anomalous events from an unsupervised approach so that labeled training data are not required. The normalcy models will also be unique for each camera, since abnormal behavior may vary by camera across a network.

In addition to detection, a computer vision benefit is the ability to automatically summarize video files and screen out irrelevant information (McCarthy and O'Mahony, 2016). One approach is to use computer vision to identify a small set of suspicious video clips in real time from multiple camera feeds or from a large video archive. A promising approach being pursued is

built on semantic indexing which uses ideas from deep learning and foreground object detection (Shah *et al.*, 2015). A temporal action localization (finding an action in long videos) approach automatically decomposes an action into several sub-actions, models each sub-action on appearance and duration into distinct steps, and detects sub-actions in an original untrimmed video. An action event usually consists of a sequence of sub-actions/sub-events in a specific order. For example, a robbery action can be decomposed into person A approaching person B, person A producing a weapon and gesturing at person B, person B holding hands aloft, handing over wallet or phone, and the two separating. The approach for localization automatically discovers the number of sub-actions for each action/event from a set of training videos, registering the point in time the action begins and ends in a video. Once identified, these segments can be flagged for human monitor review and deployment decisions.

A second video summarization approach renders a new video that highlights interesting activities in the original video and skims through redundant information to save viewing time. Along these lines, a hierarchical video summarization method is being created which will first identify small video regions termed supervoxels (regions with similar appearance and coherent motion) based on information such as color and motion. Next, high-level objects of interest such as moving humans or vehicles will be incorporated. These information sources will be combined and the defined video segments matched with previously detected and labeled objects. By detecting interesting regions as well as objects, analysis of human and object interactions is possible (e.g. a theft involving a person in a car or a fight involving multiple people).

Research question 2. Does computer vision work in the field? If a large automated camera system results in event swamping from the flagging of numerous events for review or has no significant impact on daily agency operations, computer vision's promise will be unmet. To address this issue, a set of events of interest to law enforcement and the design and installation of a computer vision workstation in a municipal police department will be evaluated. Table I lists 18 objects, events, and interactions of interest to law enforcement that computer vision algorithms are being developed to detect.

Some of the events of interest are rare and have proven difficult to locate sufficient numbers of training examples from police surveillance cameras. In addition, some events occur in conjunction with other crimes or are ambiguous. These events seldom happen without other confounding criminal activity or they are hard to identify by annotators (e.g. injured officer and custody events). Thus, it is difficult to train detectors for such less straightforward events to flag them reliably. Correspondingly, anomaly detection in the real-world assumes greater importance.

In terms of agency impact, the key computer vision field component will be a PSVAW (see Figure 3). The PSVAW will have multiple capabilities ranging from detection and localizing objects in camera feeds, labeling actions and events associated with training data, and allowing query based searches for specific events in videos. It will also be programmed to flag pre-trained criminal and new non-trained abnormal events. Using human monitor feedback, the PSVAW will refine the retrieval parameters and improve its search results over time. After repeating a number of iterations, the PSVAW will create an inductive model to detect new activity of interest in real-time so that an initial anomaly will over time become a computer vision trained, recognized, and labeled event.

Research question 3. What are computer vision impacts on a criminal justice system? The presence of surveillance cameras has been forwarded as both possibly increasing the reporting of events to the police or suppressing citizen guardianship levels (Surette, 2006). Besides issues of loss of privacy, costs, and effectiveness, because computer vision surveillance cameras are expected to catch events that humans would miss, more people may be arrested as the criminal justice net becomes wider and finer (Surette, 2005).

Interest	Operational definition
Objects	
Weapons	Handguns, rifles
Police cars	Marked law enforcement vehicles
Police officers	Uniformed law enforcement officers
Emergency vehicles	Fire trucks, ambulances (flashing emergency lights)
Left property	Abandoned bags, backpacks, etc., flagged after a set time limit
Crimes	
Criminal mischief	Graffiti, vandalism (destroying property, tipping over objects, spray painting or other property damage)
Theft	Removal of property (need to differentiate legal from illegal removal of property such as associated with property damage like breaking a car window)
Robbery	Need to differentiate legal from illegal exchanges associated with force or weapon
Burglary	Burglary from auto with window or other property damage; business burglaries constrained to "persons entering a business after closing" possibly flagging constrained from midnight to 6 AM time frame
Drug transactions	Need to differentiate legal from illegal exchanges associated with time of day or established illicit drug markets
Batteries/assaults	Shootings, stabbings, and fistfights
Public safety events	
Car crashes	Damage to moving vehicles
Crowd activity	Rioting, crowd density threshold reached, running, fighting, falling, property destruction
Individual activity	Running, falling, immobile (exceeds set time limit), blocking traffic, standing in roadway (exceeds set time limit)
Injured officer	Officer down (exceeds set time limit)
Custody events	Arrests, mental health retentions
Citizen/police	Citizen requests for assistance, officer information/identification requests
Fire/explosions	Set to minimum size limit

Table I.
Objects, crimes, and events of interest to law enforcement for computer vision algorithm development and field testing

The system-wide impact of a computer vision enhanced camera network will be assessed along two dimensions: its impact on the policing of a community through time series analysis of "all reported crime" and "calls for service" data, and the local criminal justice system's use of computer vision generated video for investigations and evidence. Utilizing measures of time spent by human monitors on video requests and processing, flagged events and response times, and use of camera video for investigations and case evidence the system-wide impact of computer vision capabilities will be evaluated.

Conclusion

Computer vision has the potential to address a wide set of problems associated with current public space camera surveillance systems from inappropriate use such as profiling and voyeurism to inherent unintended errors by human monitors. An automated camera monitor system will not view what it has not been programmed to view and when appropriately programmed will reduce the surveillance gaze from falling on unsuitable subjects. Computer vision systems can also greatly reduce the two sources of error and ineffectiveness in the use of public space surveillance cameras. A computer algorithm will not become bored or distracted during real-time monitoring so events of interest are less likely to go unseen. Simultaneously, an algorithm will not be so focused on a search for a specific event that other important events go unnoticed. These benefits are currently being developed and tested in computer vision lab settings.

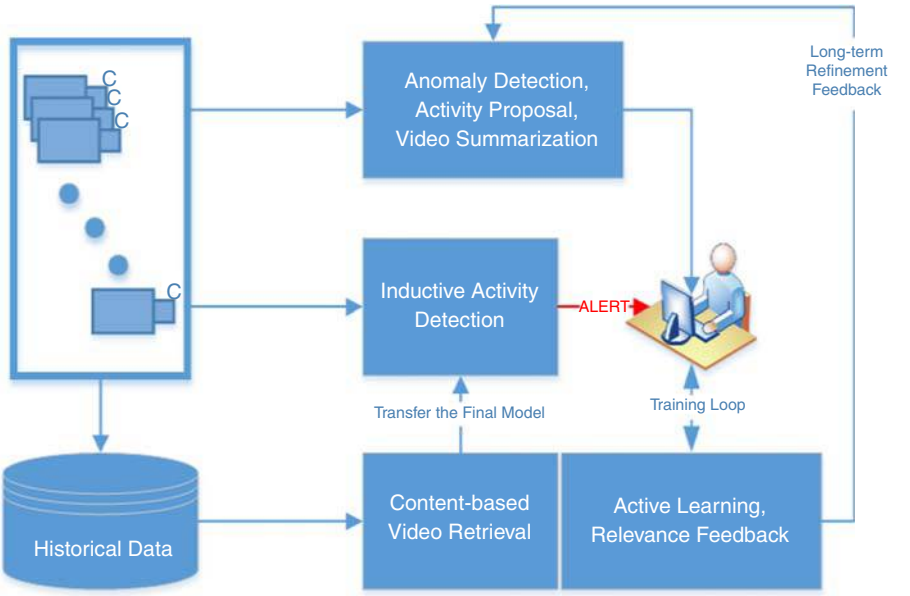


Figure 3.
Public safety visual analytics workstation (PSVAW)

Computer vision should not be expected to be a panacea for law enforcement's surveillance needs and software gaps remain such as algorithms confidently misidentifying images (Nguyen *et al.*, 2015). Additional shortfalls are due to object size (number of pixels) presenting detection errors in labeling small objects such as guns and tracking can be hampered by the occlusion of people and objects. Hence, following computer vision developments from the lab to the field will be an important step. The promise of computer vision is that the automation of monitoring can upgrade the current reality of a poorly utilized technology expenditure to a reliable public safety tool. To already budget conscious, low-on-manpower agencies a field evaluated computer vision capability stands as potentially invaluable.

Notes

1. When and how these systems work in specific applications remains under debate (Ariel, 2016; Ariel *et al.*, 2015; Williams, 2007). Indirect evidence suggests that offenders take into account the perceived level of surveillance and the likelihood of intervention when deciding whether to commit certain types of crimes, especially instrumental street crimes such as car break-ins. This suggests that easily visible cameras with signage can deter certain offenders (Gill and Loveday, 2003; Allard *et al.*, 2008; Short and Ditton, 1996; Welsh and Farrington, 2009). Spontaneous crimes such as assaults appear to be less affected and overall cameras appear most effective in reducing crime when combined with other interventions (Lim and Wilcox, 2017; Piza *et al.*, 2014a, b; Welsh and Farrington, 2004).

2. Inattentional blindness is defined as the failure to see highly visible objects directly looked at when cognitive attention is elsewhere (Mack, 2003, p. 180; Mack and Rock, 1998; see also Becklen and Cervone, 1983; Neisser, 1979; Neisser and Becklen, 1975). "Attention capture" refers to the ability of novel stimuli to gain the focus of someone otherwise cognitively engaged (Johnston *et al.*, 1990; Most *et al.*, 2005; Wolfe, 1994). The use of technology has been reported to effect both processes (Hyman *et al.*, 2009, p. 605) and inattentional blindness has been found to be a common phenomenon associated with watching video streams (Most *et al.*, 2005).

3. A classic example is demonstrated by viewers tasked with counting passes failing to see a gorilla walk through a group of people tossing a ball around. www.youtube.com/watch?v=vJG698U2Mvo
4. Over the past decade, more sophisticated alternate approaches to tracking have been presented in the computer vision literature, including those that track single and multiple objects or persons across non-overlapping multiple camera field of views (called the "hand-off" problem) and people in dense crowds (Assari *et al.*, 2016; Idrees *et al.*, 2014; Javed, Rasheed, Alatas and Shah, 2003; Javed, Rasheed, Shafique and Shah, 2003).
5. The goal in this method (termed linear discriminative analysis or LDA) is to maximize the separation of the set of images of one person from the set of images of different (but possibly similar looking on some characteristics) persons by using sets of image portraits.
6. Recent research has shown that training deep networks containing large number of hidden layers significantly improves performance on computer vision tasks such as object detection, face identification, and action recognition. However, deeper networks require larger quantities of training data compared to traditional machine learning algorithms and thus may be limited for public safety applications.

References

Allard, T., Wortley, R. and Stewart, A. (2008), "The effect of CCTV on prisoner misbehavior", *The Prison Journal*, Vol. 88 No. 3, pp. 404-422.

Ariel, B. (2016), "Do police body cameras really work", IEEE Spectrum, Posted May 4, available at: http://spectrum.ieee.org/consumer-electronics/portable-devices/do-police-body-cameras-really-work (accessed May 24, 2016).

Ariel, B., Farrar, W. and Sutherland, A. (2015), "The effect of police body-worn cameras on use of force and citizens' complaints against the police: a randomized controlled trial", *Journal of Quantitative Criminology*, Vol. 31 No. 3, pp. 509-535.

Assari, S.M., Idrees, H. and Shah, M. (2016), "Human re-identification in crowd videos using personal, social and environmental constraints", *European Conference on Computer Vision (ECCV)*, *Amsterdam*, *October 8-16*.

Baldwin, D. and Baird, J. (2001), "Discerning intentions in dynamic human action", *Trends in Cognitive Sciences*, Vol. 5 No. 4, pp. 171-178.

Barrett, H., Todd, P., Miller, G. and Blythe, P. (2005), "Accurate judgments of intention from motion cues alone: a cross-cultural study", *Evolution and Human Behavior*, Vol. 26 No. 4, pp. 313-331.

Becklen, R. and Cervone, D. (1983), "Selective looking and the noticing of unexpected events", *Memory & Cognition*, Vol. 11 No. 6, pp. 601-608.

Boksem, M., Meijman, T. and Lorist, F. (2005), "Effects of mental fatigue on attention: an ERP study", *Cognitive Brain Research*, Vol. 25 No. 1, pp. 107-116.

Bredemeier, K. and Simons, D. (2012), "Working memory and inattentional blindness", *Psychological Bulletin Review*, Vol. 19 No. 2, pp. 239-244.

Buckland, G. (2001), *Shots in The Dark: True Crime Pictures*, Little Brown & Co, Boston, MA.

Chen, B., Wang, J. and Wang, J. (2009), "A novel video summarization based on mining the story-structure and semantic relations among concept entities", *IEEE Transactions on Multimedia*, Vol. 11 No. 2, pp. 295-312.

Comaniciu, D., Ramesh, V. and Meer, P. (2000), "Real-time tracking of non-rigid objects using mean shift", *Computer Vision and Pattern Recognition*, Vol. 2, pp. 142-149.

Driver, J. (1998), "The neuropsychology of spatial attention", in Pashler, H. (Ed.), *Attention*, Taylor Francis, London, pp. 297-340.

Evangelopoulos, G., Zlatintsi, A., Skoumas, G., Rapantzikos, K., Potamianos, A., Maragos, P. and Avrithis, Y. (2009), "Video event detection and summarization using audio, visual and text saliency: acoustics, speech and signal processing", *ICASSP IEEE International Conference, Taipei, April 19-24*.

Faber, L., Maurits, N. and Lorist, M. (2012), "Mental fatigue affects visual selective attention", *PloS One*, Vol. 7 No. 10, pp. 305-315.

Fougnie, D. and Marois, R. (2007), "Executive working memory load induces inattentional blindness", *Psychonomic Bulletin and Review*, Vol. 14 No. 1, pp. 142-147.

Gao, Y., Wang, D., Yong, J. and Gu, H. (2009), "Dynamic video summarization using two level redundancy detection", *Multimedia Tools and Applications*, Vol. 42 No. 2, pp. 233-250.

Gill, M. (2003), *CCTV*, Perpetuity Press, Leicester.

Gill, M. and Loveday, K. (2003), "What do offenders think about CCTV?", *Crime Prevention and Community Safety*, Vol. 5 No. 3, pp. 17-25.

Gill, M., Spriggs, A., Allen, J., Hemming, M., Jessiman, P. and Kara, D. (2005), *Control Room Operation: Findings From Control Room Observations*, Home Office, London.

Goold, B. (2004), *CCTV & Policing*, Oxford University Press, Oxford.

Haggerty, K. and Gozso, A. (2005), "Seeing beyond the ruins: surveillance as a response to terrorist threats", *Canadian Journal of Sociology*, Vol. 30 No. 2, pp. 169-187.

Hier, S., Greenberg, J., Walby, K. and Lett, D. (2007), "Media, communication and the establishment of public camera surveillance programmes in Canada", *Media, Culture, and Society*, Vol. 29 No. 5, pp. 727-751.

Hyman, I., Boss, E., Matthew, S., Wise, B., McKenzie, M., Kira, E. and Caggiano, J. (2009), "Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone", *Applied Cognitive Psychology*, Vol. 24 No. 5, pp. 597-607.

Idrees, H., Warner, N. and Shah, M. (2014), "Tracking in dense crowds using prominence and neighborhood motion concurrence", *Image and Vision Computing*, Vol. 32 No. 1, pp. 14-26.

Javed, O., Rasheed, Z., Alatas, O. and Shah, M. (2003), "KNIGHT™: a real time surveillance system for multiple and non-overlapping cameras", *IEEE International Conference on Multimedia and Expo, Baltimore, MD, July 6-9*.

Javed, O., Rasheed, Z., Shafique, K. and Shah, M. (2003), "Tracking across multiple cameras with disjoint views", *International Conference on Computer Vision, Nice, October 13-16*.

Johnston, W., Hawley, K., Plewe, S., Elliott, J. and DeWitt, M. (1990), "Attention capture by novel stimuli", *Journal of Experimental Psychology: General*, Vol. 119 No. 4, pp. 397-411.

Keval, H. and Sasse, M. (2010), "Not the usual suspects": a study of factors reducing the effectiveness of CCTV", *Security Journal*, Vol. 23 No. 2, pp. 134-154.

Kroener, I. (2014), *CCTV: A Technology Under The Radar?*, Ashgate, Burlington, VT.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T. (2011), "HMDB: a large video database for human motion recognition", *International Conference on Computer Vision, Barcelona, November 6-13*, pp. 2556-2563.

La Vigne, N., Lowry, S., Markman, J. and Dwyer, A. (2011), *Evaluating the Use of Public Surveillance Cameras For Crime Control And Prevention*, Department of Justice, Office of Community Oriented Policing Services, Urban Institute Justice Policy Center, Washington, DC.

Levine, M., Taylor, P.J. and Best, R. (2011), "Third parties, violence, and conflict resolution the role of group size and collective action in the microregulation of violence", *Psychological Science*, Vol. 22 No. 3, pp. 406-412.

Lim, H. and Wilcox, P. (2017), "Crime-reduction effects of open-street CCTV: conditionality considerations", *Justice Quarterly*, Vol. 34 No. 4, pp. 597-626.

Lohr, S. (2012), "The age of big data", *New York Times*, February 11, pp. 1-5.

McCarthy, O. and O'Mahony, M. (2016), "End user response to an event detection and route reconstruction security system prototype for use in airports and public transport hubs", Transportation Research Board of the National Academies, Washington DC, available at: http://amonline.trb.org/trb60693-2016-1.2807374/t001-1.2823436/254-1.2823593/16-5450-1.2980693/16-5450-1.2993283?qr=1 (accessed July 1, 2017).

Mack, A. (2003), "Inattentional blindness: looking without seeing", *Current Directions in Psychological Science*, Vol. 12 No. 5, pp. 180-184.

Mack, A. and Rock, I. (1998), *Inattentional Blindness*, MIT Press, Cambridge, MA.

Marx, G. (1988), *Undercover: Police Surveillance in America*, University of California Press, Berkley, CA.

Mazaheri, A., Kalayeh, M., Idrees, H. and Shah, M. (2015), "UCF-CRCV at TRECVID 2015: Semantic Indexing", *Proceedings of TRECVID 2017*.

Memmert, D. (2006), "The effects of eye movement, age, and expertise on inattentional blindness", *Consciousness and Cognition*, Vol. 15 No. 3, pp. 620-627, doi: 10.1016/j.concog.2006.01.001. PMID 16487725.

Moeller, K. (2016), "Temporal transaction patterns in an open-air cannabis market", *Police Practice and Research*, Vol. 17 No. 1, pp. 37-50.

Most, S., Scholl, B., Clifford, E. and Simons, D. (2005), "What you see is what you set: sustained inattentional blindness and the capture of awareness", *Psychological Review*, Vol. 112 No. 1, pp. 217-242.

Näsholm, E., Rohlfing, S. and Sauer, J.D. (2014), "Pirate stealth or inattentional blindness? The effects of target relevance and sustained attention on security monitoring for experienced and naïve operators. *PloS one*", Vol. 9 No. 1, pp. 1-8, available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0086157

Neisser, U. (1979), "The control of information pickup in selective looking", in Pick, A.D. (Eds), *Perception & its Development: A Tribute to Eleanor J. Gibson*, Erlbaum, Hillsdale, NJ, pp. 201-219.

Neisser, U. and Becklen, R. (1975), "Selective looking: attending to visually specified events", *Cognitive Psychology*, Vol. 7 No. 4, pp. 480-494.

Nguyen, A., Yosinski, J. and Clune., J. (2015), "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, June 7-12*, pp. 427-436.

Norris, C. and Armstrong, G. (1999a), *The Maximum Surveillance Society: The Rise of CCTV*, Berg, Oxford.

Norris, C. and Armstrong, G. (1999b), "CCTV and the social structuring of surveillance", in Tilley, N. and Painter, K. (Eds), *Surveillance of Public Space: CCTV, Street Lighting and Crime Prevention: Crime Prevention Studies*, Vol. 10, Criminal Justice Press, Monsey, NY, pp. 157-178.

Piza, E. and Sytsma, V. (2016), "exploring the defensive actions of drug sellers in open-air markets: a systematic social observation", *Journal of Research in Crime and Delinquency*, Vol. 53 No. 1, pp. 36-65.

Piza, E., Caplan, J. and Kennedy, L. (2014a), "CCTV as a tool for early police intervention: preliminary lessons from nine case studies", *Security Journal*, Vol. 3 No. 1, pp. 247-265, available at: http://link.springer.com/article/10.1057%2Fsj.2014.17

Piza, E., Caplan, J. and Kennedy, L. (2014b), "Is the punishment more certain? An analysis of CCTV detections and enforcement", *Justice Quarterly*, Vol. 31 No. 6, pp. 1015-1043.

Piza, E., Caplan, J., Kennedy, L. and Gilchrist, A. (2015), "The effects of merging proactive CCTV monitoring with directed police patrol: a randomized controlled trial", *Journal of Experimental Criminology*, Vol. 11, pp. 43-69.

Rodriguez, M. (2010), "Cram: compact representation of actions in movies", *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference June 13, IEEE, San Francisco, CA, June 13-18*, pp. 3328-3335.

Roshtkhari, J. and Levine, M. (2013), "Online dominant and anomalous behavior detection in videos", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2611-2618.

Sampson, R. and Raudenbush, S. (1999), "Systematic social observation of public spaces: a new look at disorder in urban neighborhoods", *American Journal of Sociology*, Vol. 105 No. 3, pp. 603-651.

Sasse, A. (2010), "Not seeing the crime for the cameras?", *Communications of the ACM*, Vol. 53 No. 2, pp. 22-25.

Shah, M., Idrees, H. and Surette, R. (2015), *Studying The Impact Of Video Analytics For Pre, Live, And Post Event Analysis On Outcomes Of Criminal Justice*, University of Central Florida Center for Research on Computer Vision, Orlando, FL, Funded by US Department of Justice, NIJ-2015-R2-CX-K025.

Shah, M., Javed, O. and Shafique, K. (2007), "Automated visual surveillance in realistic scenarios", *IEEE Multimedia*, Vol. 14 No. 1, pp. 30-39.

Short, E. and Ditton, J. (1996), "Does closed circuit television prevent crime?", Monograph of the Scottish Office Central Records Unit, Edinburgh.

Soomro, K., Idrees, H. and Shah, M. (2016), "Predicting the where and what of actors and actions through online action localization", *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, June 27-30*.

St Jean, L.P. (2007), *Pockets of Crime: Broken Windows, Collective Efficacy, and The Criminal Point Of View*, University of Chicago Press, Chicago, IL.

Staples, W. (2014), *Everyday Surveillance*, Rowman & Littlefield, New York, NY.

Surette, R. (2005), "The thinking eye: pros and cons of second generation CCTV surveillance systems", *Policing: An International Journal of Police strategies & Management*, Vol. 28 No. 1, pp. 152-173.

Surette, R. (2006), "CCTV and citizen guardianship suppression: a questionable proposition", *Police Quarterly*, Vol. 9 No. 1, pp. 100-125.

Sutton, A. and Wilson, D. (2004), "Open-street CCTV in Australia: politics and expansion", *Surveillance and Society*, Vol. 2 No. 2/3, pp. 310-322.

Thomas, J. and Cook, K. (2006), "A visual analytics agenda", *IEEE Computer Graphics and Applications*, Vol. 26 No. 1, pp. 10-13.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015), "Learning spatiotemporal features with 3d convolutional networks", *IEEE International Conference on Computer Vision (ICCV), Santiago, December 7-13*.

Turk, M. and Pentland, A. (1991), "Face recognition using eigenfaces", *Computer Vision and Pattern Recognition, Maui, HI, June 3-6*.

Vlahos, J. (2009), "Surveillance society: new high-tech cameras are watching you", *Popular Mechanics,* October 1, pp. 64-69.

Wang, D., Ding, W., Lo, H., Stepinski, T., Salazar, J. and Morabito, M. (2013), "Crime hotspot mapping using the crime related factors – a spatial data mining approach", *Applied Intelligence*, Vol. 39 No. 4, pp. 772-781.

Welsh, B. and Farrington, D. (2002), *Crime Prevention Effects of Closed Circuit Television: A Systematic Review*, Vol. 252 Home Office Research Study, Home Office, London.

Welsh, B. and Farrington, D. (2004), "Evidence-based crime prevention: the effectiveness of CCTV", *Crime Prevention and Community Safety*, Vol. 6 No. 2, pp. 21-33.

Welsh, B. and Farrington, D. (2009), "Public area CCTV and crime prevention: an updated systematic review and meta-analysis", *Justice Quarterly*, Vol. 26 No. 4, pp. 716-745.

Williams, D. (2007), "Effective CCTV and the challenge of constructing legitimate suspicion using remote visual images", *Journal of Investigative Psychology and Offender Profiling*, Vol. 4 No. 2, pp. 97-107.

Wolfe, J. (1994), "Guided search 2.0: a revised model of visual search", *Psychonomic Bulletin & Review*, Vol. 1, pp. 202-238.

Ye, G., Liu, D., Wang, J. and Chang, S. (2013), "Large-scale video hashing via structure learning", *Proceedings of the IEEE International Conference on Computer Vision, Sydney, December 1-8*, pp. 2272-2279.

Yilmaz, A., Javed, O. and Shah, M. (2006), "Object tracking: a survey", *ACM Computing Surveys (CSUR)*, Vol. 38 No. 4, pp. 1-45.

Yu, C., Ward, M., Morabito, M. and Ding, W. (2011), "Crime forecasting using data mining techniques", *IEEE 11th International Conference on Data Mining Workshops*, *Vancouver, BC*, *December 11*, pp. 779-786.

Zavesky, E. and Chang, S. (2008), "CuZero: embracing the frontier of interactive visual search for informed users", *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, *Vancouver*, *October 26-31*, pp. 237-244.

About the authors

Haroon Idrees received his PhD Degree in Computer Science in 2014 from the University of Central Florida, USA and is currently working as a Postdoctoral Fellow with the Robotics Institute at the Carnegie Mellon University. His research interests include crowd analysis, object detection, action recognition, visual tracking, multi-camera and airborne surveillance, deep learning and multimedia content analysis.

Dr Mubarak Shah, Trustee Chair Professor of Computer Science, is the Founding Director of the Center for Research in Computer Vision at the University of Central Florida. His research interests include video surveillance, visual tracking, human activity recognition, visual analysis of crowd scenes, video registration, and UAV video analysis. Dr Shah is an Editor of an international book series on Video Computing; Editor-in-Chief of Machine Vision and Applications journal, and an Associate Editor of *ACM Computing Surveys* journal.

Dr Ray Surette has a Doctorate in Criminology from the Florida State University and is a Professor at in the Department of Criminal Justice, University of Central Florida, Orlando, Florida. His research interests include technology and its impact on community crime levels, media effects on perceptions of crime and justice, and criminal justice policies. He has published a number of articles and books on topics in the area of media, crime, and criminal justice. Dr Ray Surette is the corresponding author and can be contacted at: raymond.surette@ucf.edu