Law & Social Inquiry
Journal of the American Bar Foundation

Law & Social Inquiry
Volume ••, Issue ••, ••–••, •• 2018

Computer Vision and Machine Learning for Human Rights Video Analysis: Case Studies, Possibilities, Concerns, and Limitations

Jay D. Aronson

Citizen video and other publicly available footage can provide evidence of human rights violations and war crimes. The ubiquity of visual data, however, may overwhelm those faced with preserving and analyzing it. This article examines how machine learning and computer vision can be used to make sense of large volumes of video in advocacy and accountability contexts. These technologies can enhance the efficiency and effectiveness of human rights advocacy and accountability efforts, but only if human rights organizations can access the technologies themselves and learn how to use them to promote human rights. As such, computer scientists and software developers working with the human rights community must understand the context in which their products are used and act in solidarity with practitioners. By working together, practitioners and scientists can level the playing field between the human rights community and the entities that perpetrate, tolerate, or seek to cover up violations.

INTRODUCTION

In the era of social media, widespread mobile phone coverage, and the availability of the Internet to more than half the world's people, citizen media is becoming an increasingly important dimension of conflict monitoring and the documentation of war crimes, government repression, and human rights abuse. Journalists, human rights organizations, international institutions, governments, and ordinary people find themselves deluged with massive amounts of visual evidence of suffering and wrongdoing. If the documentation of the current conflict in Syria, the events of the Arab Spring in Egypt and Libya, the 2013–2014 Euromaidan Protests in Ukraine, and police violence in the United States are any indication of the future, citizen video is quickly becoming essential to our understanding of world events (Feigenson and Spiesel 2009; Sasseen 2012; New Tactics in Human Rights 2014; Wardle, Dubberley, and Brown 2014; Ristovska 2016).

Currently, manual labor by human analysts is required to extract information from conflict- and human-rights-related video. Such analysis is time consuming and, when people must be paid to do the work, prohibitively expensive. It is also

Jay D. Aronson is Associate Professor of Science, Technology, and Society at Carnegie Mellon University, where he founded and now directs the Center for Human Rights Science. He can be contacted at aronson@andrew.cmu.edu. This article was written with the support of the MacArthur Foundation (#107795), Oak Foundation (OUSA-15-025), and Humanity United (#BP10696).

emotionally challenging to repeatedly watch videos that depict horrific events like beatings, shootings, torture, suicide bombings, missile attacks, or extrajudicial killings (Dubberley, Griffin, and Bal 2015). Further, language skills, cultural knowledge, and geographic awareness can limit the number of researchers who are capable of carrying out such work, and make it difficult to enroll volunteers in analysis efforts.

While manual processing is adequate when the number of videos for analysis is small, the dissemination of conflict- and human-rights-related video in many contexts has vastly outpaced the ability of researchers to keep up. These limitations prevent human rights researchers from uncovering widely dispersed events taking place over long time periods or large geographic areas that may amount to systematic human rights violations, or force them to ignore large portions of a video collection that they know might contain useful information.

As Eyal Weizman notes, a video is most likely to be widely disseminated on the Internet or broadcast through mainstream media when it appears to show the totality of an incident "in a single image frame, both perpetrator and victim" (2017, 100). In reality, though, most eyewitness videos offer at best a partial account that must be stitched together with many other accounts to produce a more complete, but still partial, reconstruction of an event. Weizman observes:

> for every shot that includes a beater and a beaten or a shooter and a victim, there are many more that include only one or the other, or just audio, or things that happened just before or after the incident. Their relation to other images and the main incident is not obvious. It is harder to view and understand incidents that slip between images. Images containing partial information are rarely broadcast and are often discarded as trash. Searching through this image flotsam, however, we can sometimes find, synchronize, and reassemble images to reenact incidents visually and virtually in space. (Weizman 2017, 100)

Corporate, military, intelligence, and law enforcement entities have already come to the conclusion that it is crucial to automate parts of this process, and have been financing the development of machine learning and computer-vision-based systems for content capture, feature detection, behavioral analysis, and anomaly monitoring.[1] These developments have given government agencies and corporate actors an advantage over human rights practitioners, civil libertarians, and legal practitioners in the realm of video evidence.

In an effort to close this gap, various organizations and academic institutions (including one that I direct) have been facilitating partnerships between human rights practitioners and computer scientists.[2] In this article, I describe how

1. See, for example, the Intelligence Advanced Research Project Activity's ALADDIN Video Project (https://www.iarpa.gov/index.php/research-programs/aladdin-video), the FINDR Project (https://www.iarpa.gov/index.php/research-programs/finder), and the JANUS Project (https://www.iarpa.gov/index.php/research-programs/janus).

2. Some examples include the Carnegie Mellon University Center for Human Rights Science, the Columbia University/Human Rights Watch/New York University Human Rights Methodology Lab, Harvard University's Carr Center for Human Rights, SITU Research, the UC-Berkeley Human Rights Center, and the University of Cambridge's Centre of Governance and Human Rights.

computer vision and machine learning can help improve the efficiency and effectiveness of human rights practitioners who analyze video as a significant dimension of their work. As with all technologies and methods, the integration of high-throughput video analysis into the human rights domain simultaneously solves certain problems and creates others. The goal of this article is to explain why it is important to make video analysis tools available to the human rights community and to understand some potential challenges that emerge from their use.

Machine learning and computer vision offer key capabilities that manual analysis does not: first, the ability to search rapidly through large volumes of video for features or incidents of interest in the same way that one would mine a text corpus; and second, the ability to aid in the synchronization and geolocation of large event collections that lack metadata so that the relationships of the incidents portrayed can be understood better. They can therefore be used to reduce the possibility of having to rely on a single perspective at a single moment in time in investigations that have access to large volumes of video shot from multiple perspectives.

It is important to note, however, that these tools do not obviate the need to authenticate the video, nor do they provide omnipotence or a universal gaze. Event reconstruction and analysis will always be limited by the quality and completeness of available data, and human judgment is always required to verify and provide meaning and context for the work done by automated and semi-automated computing systems (Shapin 1984; Feigenson and Spiesel 2009; Landman and Carvalho 2009). Having access to more video does not get us out of this bind.

The Turn to Visual Evidence

The use of visual evidence to document conflict and human rights abuse is as old as photography itself. Archetypal images like slaves with whipping scars on their backs, Jewish concentration camp prisoners, or Nick Ut's photograph of Phan Thi Kim Phuc running down the street naked having been burnt in a US napalm attack have played a prominent role in shaping the way we think about violations of human rights.

While these images pack a powerful emotional punch, they present only a single moment in time and offer little by way of context. Thus, they must be viewed and used carefully. Law, at least, has never been a naïve consumer of visual evidence (Mnookin 1998). From the very moment that photographs were first introduced into the courtroom, legal actors had concerns. Judges, juries, and commentators held the potential of photographs to provided privileged access to the world as it actually is in tension with the power of the image to mislead the fact-finder, either intentionally or unintentionally. The fact that most judges initially analogized photography to other forms of illustrative evidence such as maps, models, and diagrams, rather than as something demonstrative and disconnected from human production, suggests that they understood—and maybe even feared—the power of the photographic image (Mnookin 1998).

Over time, however, courts came to rely on photographic images to do more than merely visually reinforce the testimony of witnesses as to matters of fact.

Photographic images were soon seen to have probative value as demonstrative evidence—that is, material evidence that could provide information to legal fact-finders above and beyond that which could be attested to by a witness. Thus, the older models, maps, and diagrams that judges once used as analogies to photography—to admit photographs into court but contain their power—ultimately achieved the status of evidence that has inherent probative value thanks to the normalization of photographic evidence.

Legal scholars have begun to examine how this turn toward visual evidence has changed legal reasoning, which has been text based for much of modern history (Porter 2014). Richard K. Sherwin (2011) tells participants in legal proceedings and investigations to pay close attention to how images and video actually form the basis for constructing reality, from both a sensory and an emotional perspective. Images and video give us the sense that we are experiencing reality, not a reconstruction, and we can often identify with being there in the moment, even though this is only a sensation produced in our mind. We need to be aware of how our brains and our culture condition the way we make meaning from images. He reminds us that videos "place viewers (like judges and juries) in the seeming position of witnesses, even though their position is limited by the camera's lens and position" and that the narrative framing and contextualization of a visual representation shapes the way it is interpreted (Sherwin 2011, 37). The playback of a video can also impact its reception—for example, by speeding it up or slowing it down, zooming in or zooming out, to alter the flow of action in a way that benefits a particular interpretation.

Sherwin's analysis raises crucial questions that we must ask as visual evidence becomes an increasingly prevalent aspect of human rights documentation and advocacy. Most importantly, we must always remain skeptical about visual claims (especially those that appear to be the most compelling), and in addition to taking note of their potentially prejudicial effects (Feigenson and Spiesel 2009), heed Ariella Azouley's reminder to ask why visual representations are produced, who produces them, the circumstances under which they are circulated, and how they are framed and contextualized (Azouley 2008).

We must also keep in mind that even with a large volume of visual evidence, facts can remain elusive. Eyal Weizman describes this problem as the "threshold of detectability" (2017, 20). This threshold occurs when a phenomenon or event in question leaves a trace on some medium, but its existence cannot explicitly be verified. The example he gives is when an object recorded on a photograph (say a person or damage to a building from a projectile strike) is close to the same size as the material element that records it (either a single salt grain in a photographic negative or a single pixel in digital imagery). This condition, he writes, "forces us to remember that the negative [or digital image] is not only an image representing reality, but that it is itself a material thing, simultaneously both representation and presence" (Weizman 2017, 20).

This article examines the possibilities, as well as the limitations, of machine learning and computer vision for human rights advocacy and accountability. In doing so, I further Weizman's call for a "counterforensics," which takes tools for which the state has a near monopoly and puts them into the hands of human rights researchers who can monitor and, when appropriate, call out the behavior of states and non-state actors alike.

Social and Legal Value

As noted above, machine learning and computer vision have the potential to find revelatory moments in large volumes of video. Perhaps more importantly, their ability to annotate large video collections and synchronize video efficiently can also limit the ability of historical revisionists or partisans to pick and choose decontextualized moments of video that obfuscate or misrepresent the past. An actor wishing to obscure the truth might show a video of protesters throwing Molotov cocktails at police without showing earlier video of police maliciously attacking the very same protesters the day before. The actions depicted in videos must be placed in context to be understood fully, and this can occur only with reasonably complete video archives that can be searched by time and place along with other, more traditional forms of forensic evidence and eyewitness testimony.

In the case of police brutality, for instance, it is important to know about previous interactions between the police unit involved and the person or people in question. Was the interaction being filmed a continuation of a set of events in which the previous events were not captured on film? Had either party issued threats that were not captured on film? Was the police unit in question normally restrained, or was the violent event captured typical of its behavior? At the same time, the ability to analyze large volumes of video does not in any way guarantee that the objective truth will be uncovered—videos, after all, still provide only a perspective on events, not an omniscient view or master narrative. At best, having access to many videos from an event, rather than a few, allows many perspectives to be shown side by side.

At minimum, video evidence also makes it much harder for violators to engage in the tactic that Stanley Cohen (2001, 7) calls "literal denial" and requires them to provide an alternative explanation for their actions or to claim that their actions were justified. The analysis of large volumes of citizen media can also help to discover and amplify alternative, community-oriented narratives that differ from those provided by large media organizations and governments.

A significant concern associated with the application of technology to human rights work more generally, however, is that the ability to mine large video collections tends to be limited to institutions with large staffs or access to expensive, technologically advanced tools and techniques. Thus, analysis of large collections of digital content quickly becomes implicated in longstanding conversations about who owns information and what can be done with it. Well-resourced human rights organizations, generally centered in North America and Western Europe, now have even greater potential to extract information from disenfranchised groups and smaller, more regional organizations and to use it to pursue their desired ends. Sometimes, these ends are at odds with, or at least not the priorities of, the groups or individuals who document conflict and human rights violations at the local level. Indeed, they often reflect the normative frameworks and imaginaries of international institutions and the individuals who staff them (Baylis 2008).

It is crucial to keep this critique of human rights documentation in mind when thinking about the tools and methods described below. One possible way to mitigate this inequity, and to ensure that human rights investigators do not cause harm by applying these tools, is to demystify them and to ensure that their power and limitations can be clearly understood by non-specialists no matter what their level of technical capacity and training. What follows is an attempt to do just this, minimizing jargon and technical nuance to provide an introduction to machine learning and computer vision that is widely accessible to the human rights community. In doing so, I lay out the basic scientific and technical underpinning of these techniques and show how they are directly applicable to the investigation of conflict and human rights violations.

MACHINE LEARNING AND COMPUTER VISION FOR EVENT DETECTION

Machine Learning

Machine learning is the science of creating computing systems that are programmed to arrive at logical conclusions about the world through exposure to, and processing of, data (Geitgey 2014). Machine learning uses algorithms—a set of processes or rules that are performed in a step-by-step manner—to enable systems to sort dynamically through a corpus of data to discover regularities and relationships. In machine learning, computer scientists do not hard code rules of action that a system uses to do the exact same thing every time; rather, they give the system the capacity to recognize patterns, trends, or categories in “training” data provided by the programmer and then to use those discoveries to analyze another set of data that has not yet been processed.

Machine learning systems sometimes use supervised training, in which the programmer gives the system a labeled data set consisting of the thing of interest, and asks the computer to figure out how to recognize similar things in other data sets that it has not seen before. One example would be giving a machine learning system thousands of sample sentences with nouns, verbs, adverbs, and adjectives labeled, and then asking the algorithm to label these entities in sentences to which it had never previously been exposed. The programmer does not tell the system that a word ending in “ing” is usually a verb or that capitalized words appearing anywhere other than the first position of a sentence are generally nouns; instead, she lets the system figure that out based on the labels in the training data. It is important to note that the effectiveness of a supervised machine learning model is directly related to the quality and representativeness of the training data provided to it. The system will pick up on patterns in the data, and it will therefore retain any biases embedded within those data. For instance, if a machine learning system is never exposed to words ending in “ing” that are nouns or adjectives (e.g., king or interesting), but is exposed to many verbs ending in “ing,” it may come to the very logical (but incorrect) conclusion that words ending in “ing” are verbs even when they do not behave like most other verbs. The system can obviously also be trained to recognize nouns by their position in a sentence, but this must be a conscious choice of the programmer.

Scientists can also use unsupervised training, in which the system is provided with an unlabeled data set and is programmed to identify the strongest categories or structuring principles (often called clusters) within it. An example of unsupervised learning would be giving a machine learning system demographic and life history information about thousands of people who receive PhDs in a particular discipline (or, alternatively, have been convicted of a particular crime) and asking it to determine which factor or factors seem most predictive of this outcome. It is important to note that the characteristics of the data set determine what can be learned even if it has no labels. If the PhD data set is made up primarily of humanities professors rather than computer scientists, or if the criminal data set is made up primarily of individuals convicted of possession or sale of methamphetamine rather than a broader set of drug crimes, the system will come to certain conclusions about that particular population that may not be true of the broader population.

Regardless of whether supervised or unsupervised training is used, the output of the machine learning system's analysis becomes a set of "classifiers" that can then be applied to other data sets that have not been used in the training process. Classifiers are reductive—they take the complexity of the world and convert it into decision processes that can be used to identify similar things in other contexts. They are rarely 100 percent accurate, and they can lead users astray if their limitations are not understood and taken into account. For instance, a language processing classifier trained on newspaper articles and official government documents may not do a great job of analyzing transcripts of conversations in slang or regional dialects. One would require training data with these variations to create classifiers that would work well on them. Ideally, machine learning systems should be trained on diverse material to ensure that they are not overspecialized for a single context.

Computer Vision

Computer vision is the analysis of digital visual images to understand both the objects that are depicted within them and the scenes from which they were constructed. Computer vision can involve detection of specific objects, segmentation (separation) of multiple objects within one scene, tracking these objects over time and space, three-dimensional reconstruction of the objects and/or scenes, and determination of the placement of the camera in the scene over time and/or space. Numerous methods are used to carry out these tasks, including color analysis, shadow and illumination analysis, geometrical analysis of curves and edges, and photogrammetry (mathematical and geometric techniques to make measurements within the image) (Szeliski 2011).

When combined with machine learning, the principles of computer vision can be used to identify objects in, and reconstruct scenes from, digital images. One simple but illustrative example of computer vision is teaching an algorithm to recognize letters or numbers. We know from experience that both typography and human handwriting vary widely, but we learn to recognize even the most challenging writing based on our general understanding of what letters and numbers look like, as well as our contextual knowledge of language and mathematical representation. Even if one or a few individual characters are indecipherable, we can generally infer what words or ideas the writer is representing.

Machine learning systems can be trained to read a huge range of typography and handwriting with the aid of computer vision. The situation becomes more complicated when the letters or numerals are not located on a solid, white background; they may be integrated within a visual scene such as a photograph or painting, be presented at low resolution, or be distorted (as most of us know from solving the Captcha and ReCaptcha puzzles that confirm our identities as humans). Distortion represents a significant computational challenge, especially when the object of interest is more complex than a letter or numeral—such as evidence of a possible human rights violation—and scientists devote a great deal of effort to developing tools to enhance the resolution and clarity of objects within an image.

APPLYING MACHINE LEARNING AND COMPUTER VISION

In this part, I discuss video analysis work that is representative of the current and future potential of machine learning and computer vision for detecting human rights violations. I also discuss potential barriers to widespread dissemination of these technologies within the human rights community, the impact that these technologies might have on human rights advocacy and accountability efforts, and ethical concerns that emerge from their use.

Event Detection: E-LAMP

Machine learning systems can learn to detect many of the kinds of objects that are of interest to human rights researchers, including tanks, missiles, helicopters, airplanes, military vehicles, particular styles of building, soldiers in camouflage, large crowds, corpses, and visually distinct geographic locations like bridges over water, mountainous terrain, or a desert (Aronson, Xu, and Hauptmann 2015). They can also detect and tag sounds like gunshots, explosions, cries, or screams (Liang et al. 2016). This allows a researcher or organization with limited human resources or time to scan a large video collection rapidly to find material relevant to an investigation. This makes it less likely that the investigator will have to rely on a single video to understand an event and also increases the amount of data that can be used to place the event of interest in context.

Event detection is computationally complex because it involves the detection of numerous semantic concepts (i.e., objects, sounds, scenes, and actions) taking place in a dynamic environment. While this task is challenging enough when carried out with video from a fixed camera in a constrained environment, the kinds of videos relevant to human rights investigations are significantly more complex. Typically, the videographer (often an amateur) is in motion and reacting to the events he or she is capturing on camera. The camera thus is unconstrained in time and space and generally offers something akin to first-person vision (Tong et al. 2014). Material drawn from social media contains an almost infinite number of situations that are not bound by clear-cut rules or spatial environments (Tong et al. 2014). Further, social media videos capture an almost infinite variety of human and non-human behavior. To mine this resource effectively, an analyst needs to be able to develop novel classifiers on the fly.

One system, developed by researchers at Carnegie Mellon University, that directly addresses these challenges is called "Event Labeling through Analytic Media Processing" (E-LAMP). It works at the most basic level when an operator provides the system with a set of training videos or video shots that depict a particular activity or thing along with a set of null videos that depict other unrelated activities. E-LAMP analyzes these videos for a variety of the kinds of features described above (selected in various combinations based on the need for accuracy, speed, and processing capacity), which can be combined into a computational machine learning model of the relevant action or event. The system then delves into the larger collection of videos to look for other potential examples of this model. It returns a set of videos to the operator that it thinks match the activity in question. The operator confirms whether the proposed matches are correct or incorrect, and E-LAMP takes this information into account and tries again. Once the system returns mostly correct results (which are rarely 100 percent accurate for a variety of reasons), this set of patterns is labeled as a classifier, or event kit, for the particular action (Tong et al. 2014).

This classifier, which can be visual, aural, semantic, or a combination of the three, can then be used to search for particular instances of it in any other video collection. The classifier may need to be modified to work well in these other contexts. In addition to this kind of search capability, E-LAMP can also be used to detect duplicates or near-duplicates, which is valuable in human rights contexts because it is possible to see how similar footage is used in different videos and how various users edit and recombine footage. It also makes it possible to gain multiple perspectives on a single event.

In addition to visual classifiers, E-LAMP can currently perform speech recognition in English and Arabic (together with preliminary versions of Spanish, German, French, Mandarin Chinese, Cantonese, Turkish, Pashto, and Tagalog), and can detect a variety of unique sounds, such as gunfire, explosions, airplanes, and helicopters. Text found in videos, whether as subtitles, titles, or signs, is processed through optical character recognition (OCR) and rendered searchable, although researchers are still working to improve the accuracy of OCR in the kinds of low-resolution contexts found in video.

E-LAMP functions by processing the video once in its entirety, segmenting it by scene, and then taking a visual snapshot (a "keyframe") at the midpoint of each segment. The feature detection described above is then performed on these frames so that they can be compared with classifiers for the targeted event. The results are returned with probability calculations—the higher the score, the more confident the system is in the call. In most cases analyzed, the higher probability results tend to be accurate, while no systematic claim can be made about the lower probability results. In some cases, they are still accurate, and in other cases they are not (see below). Currently, eight-hundred hours of video would cost approximately $1,872 for computing resources plus 180–200 GB of storage space.[3]

Building Classifiers

To provide a first test of the capacity of the system to aid in conflict monitoring and human rights investigations, Carnegie Mellon researchers sought to identify certain categories of weapons depicted in a set of approximately five-hundred videos that focused on events taking place in Aleppo, Syria, in late 2013. Given the nature of the conflict in Syria, rebel groups routinely filmed their military exploits and regularly reported about their caches of weapons. It is, of course, important to recognize that all such videos are public relations ploys, and that groups avoid distributing footage that highlights their weaknesses or shows them being defeated. One cannot make statistical calculations of any sort based on available social media reports because they are only a convenience sample of data (Price, Gohdes, and Ball 2015). That said, one can still gather quite a bit of general information about access and availability of weapons systems from these videos—particularly the presence of weapons systems in a particular region.

At the moment, there are over twenty-five hundred preexisting semantic concepts available in E-LAMP for use in searching a video collection. Because the system was not trained for conflict and human rights analysis, though, classifiers for much of the material of interest to researchers in this domain will have to be built from scratch. Thus, when researchers first uploaded the Aleppo video collection, there was no semantic concept for "mortar launcher," or for most other weapons systems. There were generic classifiers already built for "weapon" and "machine gun" (both of which were likely developed from video collections that included only recreational shooting, hunting, or gun collecting), but this would be useful for only a fraction of the weapons used in the Syrian conflict.

To try to get around this "cold start" problem, the operator can begin with an analysis of available metadata (which only works with video with at least some metadata preserved), speech recognition, or OCR words (in this case Arabic), or try her luck with the available classifiers. Thus, if the Arabic words for "mortar launcher" are written or spoken in any of the videos in our collection or in their metadata, they will at least theoretically be detected in response to the user's query even if the system has never been trained to recognize an actual mortar launcher. The researcher can also manually select a few videos of mortal launchers that she knows are in the larger video collection.

Assuming that the investigator has a few videos containing mortar launchers, she then can either play the videos in their entirety and select particular "shots" of interest (e.g., moments when the mortar launcher is visible or heard being launched), or use keyframes to select a few positive videos shots containing mortar launchers (Figure 1). These shots and/or keyframes are then fed into the machine learning model. The model will utilize audio, image, and motion signatures

3. Personal communication, Alex Hauptmann (Research Professor, Carnegie Mellon University School of Computer Science), January 2017.

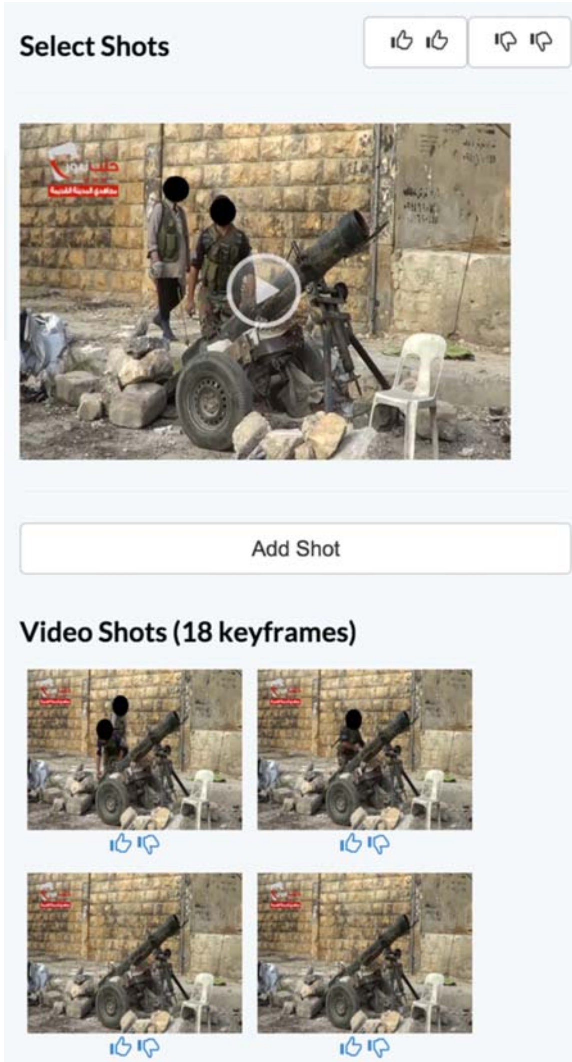

FIGURE 1.
Initial Mortar Launcher Keyframes from the First Video We Found that Contained this Weapon System
***Notes*: Eighteen total keyframes were taken by E-LAMP and we selected a few of these to build the model for our classifier. We also selected keyframes from other videos to broaden the variety of mortar launchers detected and ensure that the angle, positioning, background, and type did not overly limit the machine's final model. [Color figure can be viewed at wileyonlinelibrary.com]**

extracted from these video shots to create a classifier that will become the basis for a new semantic concept of "mortar launcher" that can be refined through a variety of mechanisms until satisfactory results are achieved in a test data set. In the case of the mortar launcher classifier, for example, the model initially misidentified things like power lines, truck mounted anti-aircraft guns, and other linear objects with similar backgrounds as positive examples.

The operator has to determine the ideal sensitivity of the classifier for her purposes. A very sensitive classifier will have a high degree of accuracy in the high confidence range, but will miss many positive but lower confidence cases. A less sensitive classifier will capture a greater percentage of positive video cases (both high and low confidence), but will also capture many incorrect ones as well. This means more time needs to be spent by the operator separating the signal from the noise. Once built, the model can be applied to the entire video collection or any other appropriately processed set of videos, although accuracy may decrease when applied to a new collection.

E-LAMP has also been tested on a larger collection of 13,570 publicly available YouTube videos with the goal of identifying more complicated visual phenomena, including helicopters (which were routinely used in the Syrian conflict by the Assad regime to drop barrel bombs on neighborhoods held by antigovernment groups) and corpses, using the process described above. Both were successful in identifying the object in question with a high degree of accuracy in the top results. The most common misidentification for the helicopter classifier was an airplane, although there also appear to be a few incidental images from the scraping process (including a pirate flag with skull and bones that appear to mimic the rotors of a helicopter and a gecko mascot that appears to be falling from the sky with its four limbs spread out in a US insurance company advertisement) that show up toward the bottom of the top 100 results (Figure 2). The "corpse" (defined by Carnegie Mellon researchers as bloodied bodies with visible faces in a horizontal pose with no movement) detector was similarly successful. On the first iteration, ninety-five of the top one-hundred hits were correct. The computational model mistook what appears to be an open artichoke flower for a corpse because of the similarity of the shape and contrast to the face of a corpse, as well as a couple of images of what appear to be pink blossoms against dark green leaves (search results available upon request). It is important to note that this classifier likely missed many cases of corpses with unexposed faces, or those lacking blood, so a separate one would have to be built for those cases.

Carnegie Mellon researchers are currently working with a variety of human rights partners to determine how E-LAMP can be integrated into their organizational workflows. So far, E-LAMP seems to be most useful in acting as a filter to remove irrelevant videos from the analysts' work queue and pinpointing where particular entities of interest are within a collection of videos. The technology has not advanced to the point where it can be relied on to tag a large video collection and populate a database with the results for investigative and analytical purposes automatically without significant human cross-checking.[4]

Face Detection and Recognition

Face detection and recognition is a particular application of machine learning and computer vision. Over the past decade, computer scientists have developed

4. Personal communication, Alex Hauptmann, January 2017.

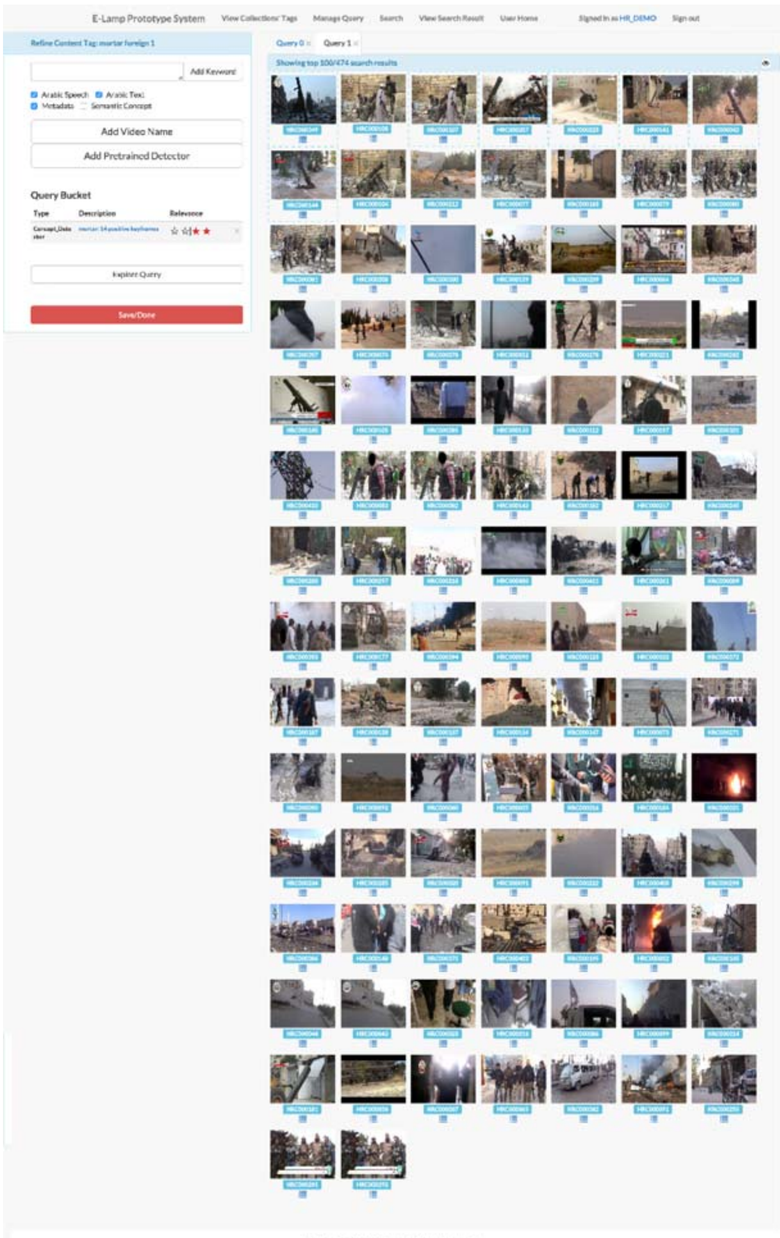

FIGURE 2.
Top 100 Mortar Launcher Videos Out of 476 total. [Color figure can be viewed at wileyonlinelibrary.com]

tools that make face detection relatively routine. This should not be particularly surprising: although all human faces are unique, barring unusual circumstances, they share certain common anatomical landmarks that are related to one another in a predictable fashion—two ears at the side of the face and in the center two eyes that sit above the nose, which sits above the lips, which sit above the jaw.

Detecting partially occluded faces or faces turned to the side is more challenging, but computer scientists have developed sound methods for solving this problem.

A much more complicated problem is recognition: determining whether faces from different images belong to the same person. If the two images were taken straight on in the same lighting, from the same distance, with the same expression, and in high resolution, then the problem can be solved by measuring enough facial features and knowing how many one needs to combine in order to achieve the desired level of certainty that the match is not a false positive. These conditions might hold for mug shots or driver's license/passport photos, which are taken under standardized conditions, but in real-life situations, they usually do not. As a result, computer scientists have had to develop methods to compare faces in images that differ in terms of lighting, angle, pose, facial expression, composition, resolution, distance, and other characteristics. Most importantly, before any pair-wise comparisons can take place, the faces need to be computationally "transformed" so that the eyes, nose, and mouth are centered and are presented in the same scale, with the same expression, all without introducing any significant distortion or error. It goes without saying that this is a difficult task, and it can lead to problems in lower-resolution images (Geitgey 2014).

Because geometric feature-based methods can be cumbersome (especially after the kind of transformation just described), scientists have developed photometric techniques that rely on differences at the digital level rather than analysis of physical features. This makes it easier for algorithms to match faces because they can be treated as data rather than particular physical objects. The first successful method compares the intensity of light of each pixel to its neighboring pixel and records the direction from the lighter pixel to the darker pixel. Once these relationships are summarized using various computational techniques, faces can be matched by comparing gradients and their interrelationships. More recent methods use convolutional neural networks to do similar kinds of computations, but more rapidly and accurately (Geitgey 2014). In addition to photometric methods, computer vision experts are also developing methods to reconstruct faces in three dimensions by combining two-dimensional data extracted from multiple frames showing the face from a variety of perspectives. These methods are not yet well understood or ready for widespread use, but the intelligence community and private companies are actively developing them.[5]

In theory, facial recognition has a plethora of potential uses in the human rights domain. Putting aside massive ethical concerns for a moment, and not focusing on the negative applications of the very same technology by states or powerful non-state actors, one could imagine human rights investigators using it to identify perpetrators and victims of a filmed violation, or even bystanders who could then be contacted to provide important information or testimony in an investigation (Georgetown Law Center on Privacy and Technology 2016).

5. See, for example, the Intelligence Advanced Research Project Activity's JANUS Project (https://www.iarpa.gov/index.php/research-programs/janus).

In practice, however, there are technical limitations that make its use in human rights contexts less likely in the near future. Facial recognition is becoming increasingly more accurate from high-resolution photographs and high-resolution surveillance video (which do occasionally become relevant in human rights investigations), but the low resolution and highly processed nature of most videos recovered from social media do not generate enough fine-grained data for facial recognition systems to measure enough characteristics of the face to generate meaningful matches.[6] Such tools may be used to reduce the number of faces that need to be manually checked for a match, but they cannot provide accurate matches. Although methods of image enhancement are in development, they are not yet ready for use.

Along these lines, many available facial recognition systems do a poor job of identifying people of African ancestry compared to Asian or European descendants. Why? Most likely because developers of these systems are of Asian or European descent and, intentionally or unintentionally, tend to use training data sets that do not include people of African descent (Phillips et al. 2011; Garvie and Frankle 2016; Orcutt 2016). Over the next few years, facial recognition systems will undoubtedly become better at recognizing faces from around the world, spurred by bad press surrounding their limitations and the likelihood that corporate, intelligence, and military actors will pay for such capabilities.

Second, in many videos related to human rights situations, faces are obscured with adornments like thick beards and head coverings, leaving only a small portion of the face available for scrutiny. There are research groups building algorithms to recognize a face based only on a small visible portion, but such systems are not currently reliable enough for widespread use (Juefei-Xu, Lu, and Savvides 2015). At the same time, especially when dealing with casualties of war or human rights violations, faces of victims (whether alive or deceased) are often different from the way they look in their last available photograph. Damage from blunt force trauma, drowning, burns, starvation, desiccation, and other factors significantly alter the characteristics of the face to the point that is difficult for even a human investigator to recognize a match. Even visual recognition by family members becomes difficult when bodies are not stored under controlled conditions in hot, moist climates (Morgan et al. 2006). The tools available to model what the person's face might have looked like pre-trauma are not reliable enough to use in most investigations except to provide general leads.

Event Reconstruction Though Multi-Perspectival Video

Computer scientists are increasingly interested in using machine learning and computer vision to reconstruct events in time and space. This provides a visual narrative that can help fact-finders to understand better what happened during an event, as well as the context in which the event took place (Forensic Architecture and SITU Research 2010, n.d.; Weizman 2017). Early work in this field, most

6. Personal communication, Alex Hauptmann, January 2017.

notably Project Rashamon at UC-Berkeley, focused on developing tools for simultaneously displaying several synchronized videos of an event in a way that made it possible to see the event from multiple perspectives at the same time (Lafrance 2014). Although this tool was an important first step, it was constrained by the need to have timestamps for synchronization, which are generally stripped during upload to social media, and the fact that the system presents several video feeds to the viewer, which can become overwhelming quite quickly.

Recent work in event reconstruction has focused on bringing together multiple streams of data into a single, coherent account that flows through time and space. Forensic Architecture and SITU Research pioneered this work in the human rights context through their reports on human rights violations in Israel/Palestine in collaboration with various local and international human rights practitioners. In describing its approach to event reconstruction in a report on a series of attacks on the city of Rafah in response to the kidnapping of an Israeli soldier by Palestinian forces in August 2014, Forensic Architecture notes that its partner organization, Amnesty International, did not have direct access to Gaza to conduct a fact-finding mission. It was, however, able to collect eyewitness testimony and photographs and videos shot by citizen witnesses and professional journalists remotely though social media and local partners. Amnesty also obtained high-definition satellite imagery from the Pléiades satellite. Given this evidence base, Forensic Architecture developed methods of superimposing visual evidence onto a 3D model of Rafah constructed by parametric modeling of landmarks in photography-based visual surveys of the area and information from the satellite imagery (Amnesty International and Forensic Architecture n.d.)·

Once visual evidence could be placed geographically, Forensic Architecture synchronized all available visual evidence to a universal clock. This was done by cross-referencing eyewitness testimony, available metadata, and visual information, such as smoke, contained within the images and videos. It also undertook shadow analysis to corroborate time and place calculations. By mapping evidence over time and space, analysts and fact-finders can "explore the spatial and temporal connections between the various sources and reconstruct the unfolding of events" (Forensic Architecture n.d.). This analysis can be conducted at the urban scale (e.g., by tracking movements of groups of people engaged in protest or conflict over geographic space and time) or at the personal scale (e.g., capturing multiple views of a person being hit by a projectile, reconstructing its path from the weapon to impact, and comparing visual evidence to autopsy reports or eyewitness testimony) (Forensic Architecture and SITU Research 2010 n.d.).

Semiautonomous Synchronization and Geolocation

To put together a video archive that is searchable over time and space and that can be mapped onto the location where it occurred with relative accuracy, one must synchronize and geolocate relevant videos. The methods used by Forensic Architecture in its Rafah report are completely manual. This is fine if one has limited video, numerous skilled human analysts, and/or no time deadline. In today's world of widely distributed mobile phones, limited resources, and pressing deadlines, these conditions are rarely met.

In one recent example, Ukrainian human rights lawyers representing families of people killed or injured by national police during the Euromaidan protests in Kiev in 2013 and 2014 were overwhelmed by the dozens of hours of videos that were relevant to their cases. They needed to find evidence to support their claim that security force personnel shot these individuals despite the fact that they were not immediate threats to police or other bystanders. Because most of this material was gathered from social media, it lacked technical metadata. Some video, which was received from news outlets or other official sources, did have reliable metadata and could serve as a basis upon which to synchronize the video lacking metadata.

The legal team initially asked an analyst who was familiar with the events of Euromaidan and the geography of Kiev for help in synchronizing their large video collection. Over the course of eight months, this individual managed to stitch together a small percentage of the total video using visual and audio cues in the recordings and the few videos with reliable time stamps. Because videographers tended to be relatively far away from one another and focused on different points within a larger scene, and because the scenes depicted in the videos tended to be smoky and monochromatic, this task was incredibly challenging and tedious. Often, there are no visual cues that clearly linked two synchronous videos. The analyst put this work together in a nine-channel video grid that stretches 4h37m19s, with large gaps throughout. This was an impressive feat given the circumstances, but it was clear both to the lawyers and the analyst that much of the video could never be synched using manual methods.

The legal team eventually requested the assistance of the computer scientists who created the E-LAMP system described above to develop semiautonomous methods for synchronizing video over time and geolocating the camera. Because visual cues were sparse in these videos and there was no metadata to rely on, the researchers decided to use the kinds of audio signal processing described above. They developed an algorithm that recognized a standardized vocabulary of "features" (e.g., wind, screaming, gunshots, airplane noise, amplified speeches, music, and explosions) and then compared the sequence of these features in each clip to all others in order to look for reasonable matches (see Figure 3).[7] It was necessary to include a certain amount of tolerance for imprecision because each recording device captures sound in a slightly different way due to the specificities of the microphone, the distance from the sound source, and the other noises present in the immediate vicinity of the videographer.

It is once again important to note that machine learning algorithms are only probabilistic—to ensure accuracy, all synchronizations need to be checked manually by a human analyst. This analyst also needs to adjust the offset that results from videos being shot at different distances from the source of the sound used for synchronization (sound travels more slowly than light). All told, the Carnegie Mellon team was able to synchronize 4h16m13s, solely through their algorithms, some of which overlapped with the existing 4h37m19s, synchronization. The combined total

7. Personal communication, Alex Hauptmann, January 2017.

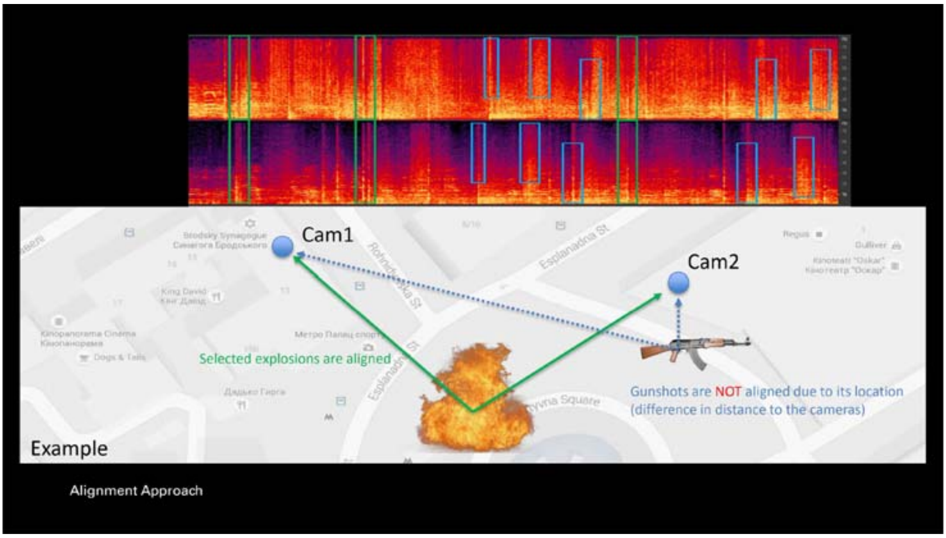

FIGURE 3.
Audio-Based Synchronization [Correction added on XX March 2018 after first online publication: the figure caption has been corrected from "Audio-Based Synchronization Pipeline" to "Audio-Based Synchronization".]. [Color figure can be viewed at wileyonlinelibrary.com]

of synchronized video was increased approximately 50 percent, to 6h52m10s, through these efforts. The scientists also determined that 10h40m02s, of the 52h24m00s, of the video analyzed were duplicates of other video in the collection. Once the system was developed, the synchronization took a few days in total rather than eight months.

To aid in the geolocation process, the research team developed an algorithm that used computer vision principles to compare the background scene in a video to a library of accurately geotagged images that were mined from sources such as Google Street View, Flickr, or images taken specifically for this task. Because the results of this system are probabilistic, they also created a tool to allow a human analyst quickly to confirm or reject a match between the scene in a video and a geotagged image. If confirmed, the system quickly moves on to the next video. If rejected, the next most likely location image shows up and the human analyst can repeat the process.

Unfortunately, the utility of this system is dependent on the quality of the videos under review and the availability of accurately geotagged images for comparison. In this first test, the algorithms were only able to geolocate approximately 12 percent of the videos in question. They also determined that 21 percent contained no information that could be used for geolocation (either because they were shot indoors or because no physical landmarks were visible).[8] These results can be improved with refinement of the algorithms developed, but they will likely never replace human knowledge and judgment.

8. Personal communication, Alex Hauptmann, January 2017.

CONCLUSION

Machine learning and computer vision make it possible to gather more information about an event, but they are still limited to available data. They will not magically correct for biased or incomplete data in a data set, but they can make it possible to discover content that may otherwise be obscured by sheer volume. Nor will these technologies allow investigators or analysts to tell the story of an event or present an issue "as it really is." Video evidence, like all forms of evidence, is inherently perspectival, and it is almost always biased in some way. Some crimes or abuses will go unrecorded because the victims or witnesses are unable or unwilling to film them, or because perpetrators or other actors seize the recordings before human rights advocates preserve them. There is by now a well-developed literature on dangers of disappearing content, the role that social media platforms and other companies can play in saving these data, and other issues of video preservation (Aronson 2017).

The fact that human verification of the outputs of machine learning and computer vision systems is required to ensure their validity will likely blunt some of the legal challenges that may emerge when they are used in justice and accountability contexts, including domestic courts, international tribunals, or truth commissions. In this sense, these systems are not producing evidence, but are instead narrowing down the amount of material that must be examined by human analysts who are making factual claims that they then present in justice, accountability, or advocacy settings.

Concerns about the validity of the output of machine learning systems can be further addressed by ensuring that the code that goes into them is open source and available for review and testing by outside experts. These systems gain credibility by being constructed with code that is subject to the scrutiny both of users and those who find themselves involved in investigations (Wexler 2015a, 2015b).

The use of technologies like machine learning and computer vision can reinforce or exacerbate existing inequalities in the human rights domain (Piracés 2018). They require a level of expertise and computing resources that put them out of the reach of many human rights advocates, particularly those outside of North American and Western Europe or those without connections to academic research centers. Several organizations are working to ensure that the technologies described in this article are accessible to as broad a swath of the human rights community as possible by making them as user friendly as possible and compatible with cloud services like Amazon Web Services, which provide on-demand storage and processing at reasonable costs. At the moment, though, the use of these tools is limited to organizations that have generally high levels of technical capacity, even if they are not particularly well-resourced.

The technologies described in this article are likely already being used by many governments that violate human rights. The ability to rapidly scan large volumes of video of protest for faces that can be manually identified, or to search automatically for written or spoken names in thousands of videos, or to identify the precise location where a video of dissidents or protesters has been filmed can cause real and direct harm to individuals documenting human rights violations and the activists and ordinary people they film. Unlike weapons systems, which require specific material or equipment that can be regulated by suppliers, there is little the computer science community can do to regulate the distribution and use of these computer vision and machine learning technologies. This problem is sharpened by the fact that open source and easily accessible code are crucial to their widespread dissemination beyond a very narrow and elite sliver of the global human rights community.[9] There are no easy fixes for this problem. What computer scientists can do is ensure that the human rights community has equal access to the tools of computer vision and machine learning so that human rights practitioners can return the gaze of violators and engage in counterforensics (Weizman 2017).

Counterforensics requires close partnerships between technology developers and human rights practitioners (Piracés 2018). Both have crucial roles to play. Human rights violations are too complex and too context dependent to be discoverable solely by an automated system without significant input of prior human knowledge and verification of the outputs of computer systems. The integration of computing technology into human rights work requires knowledgeable practitioners who understand the legal and evidentiary requirements of advocacy and accountability efforts and the ultimate objectives of the human rights community.

At the same time, technologists need to work closely with practitioners to ensure that they do not place new and unrealistic technical or resource demands on organizations that adopt new technologies, or promote unsustainable dependencies on outside expertise. Finally, it is important to remember that the technological systems described in this article are best thought of as tools—they cannot document human rights violations on their own, but can increase the efficiency and effectiveness of human analysts when used properly.

REFERENCES

Amnesty International and Forensic Architecture. "'Black Friday': Carnage in Rafah, 'Methodology'." n.d. https://blackfriday.amnesty.org/methodology.php (accessed February 15, 2017).

Aronson, Jay D. "Preserving Human Rights Media for Justice, Accountability, and Historical Clarification." *Genocide Studies and Prevention* 11, no. 1 (2017): 82–99.

Aronson, Jay D., Shicheng Xu, and Alex Hauptmann. "Video Analytics for Conflict Monitoring and Human Rights Documentation." 2015. https://www.cmu.edu/chrs/documents/ELAMP-Technical-Report.pdf (accessed February 15, 2017).

Azouley, Ariella. *The Civil Contract of Photography*. Cambridge, MA: MIT Press, 2008.

Baylis, Elena. "Tribunal-Hopping with the Post-Conflict Justice Junkies." *Oregon Review of International Law* 10 (2008): 361–90.

Cohen, Stanley. *States of Denial: Knowing About Atrocities and Suffering*. Cambridge: Polity Press, 2001.

Dubberley, Sam, Elizabeth Griffin, and Haluk Mert Bal. "Making Secondary Trauma a Primary Issue: A Study of Eyewitness Media and Vicarious Trauma on the Digital Frontline." 2015. http://eyewitnessmediahub.com/research/vicarious-trauma (accessed February 15, 2017).

---

9. For more on dual-use technologies, see the Wassenaar Agreement, which regulates the transfer of technologies that can be used for terrorism or other destabilizing activities by participating states: http://www.wassenaar.org/.

Feigenson, Neal, and Christina Spiesel. *Law on Display: The Digital Transformation of Legal Persuasion and Judgment* New York: New York University Press, 2009.

Forensic Architecture. “Rafah: Black Friday.” n.d. http://www.forensic-architecture.org/case/rafah-black-friday/#toggle-id-4 (accessed February 15, 2017).

Forensic Architecture and SITU Research. “Report: Summary of Findings on the April 17, 2009 Death of Bassem Ibrahim Abu Rahma, Bil’in.” 2010. http://www.situstudio.com/blog/wp-content/uploads/2010/07/Abu_Rahma_report_web.pdf (accessed February 15, 2017).

——. “The Use of White Phosphorus Munitions in Urban Environments” n.d. http://www.situresearch.com/white_phosphorus/index.html.

Garvie, Clare, and Jonathan Frankle. “Facial-Recognition Software Might Have a Racial Bias Problem.” *The Atlantic*, April 6, 2016. http://www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/ (accessed February 15, 2017).

Geitgey, Adam. “Machine Learning Is Fun! The World’s Easiest Introduction to Machine Learning.” 2014. https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471#.hxc2prhlu (accessed February 15, 2017).

Georgetown Law Center on Privacy and Technology. *The Perpetual Line-Up: Unregulated Police Facial Recognition in America*. Washington, DC: Georgetown University Press, 2016.

Juefei-Xu, Felix, Khoa Luu, and Marios Savvides. “Spartans: Single-Sample Periocular-Based Alignment-Robust Recognition Technique Applied to Non-Frontal Scenarios.” *IEEE Transactions on Image Processing* 24, no. 12 (2015): 4780–95.

Lafrance, Adrienne. “Idea: Syncing Multiple Videos from the Same Event on YouTube.” *Atlantic*, 2014. http://www.theatlantic.com/technology/archive/2014/04/hey-youtube-we-need-a-multi-video-syncing-function/361236/ (accessed February 15, 2017).

Landman, Todd, and Edzia Carvalho. *Measuring Human Rights*. London: Routledge, 2009.

Liang, Junwei, Susanne Burger, Alex Hauptmann, and Jay D. Aronson. “Video Synchronization and Sound Search for Human Rights Documentation and Conflict Monitoring.” 2016. http://www.cmu.edu/chrs/documents/Video-Synchronization-Technical-Report.pdf (accessed February 15, 2017).

Mnookin, Jennifer L. “The Image of Truth: Photographic Evidence and the Power of Analogy.” *Yale Journal of Law & the Humanities* 10, no. 1 (1998): 1–74.

Morgan, Oliver W., Pongruk Sribanditmongkol, Clifford Perera, Yeddi Sulasmi, Dana van Alphen, and Egbert Sondorp. “Mass Fatality Management Following the South Asian Tsunami Disaster: Case Studies in Thailand, Indonesia, and Sri Lanka.” *PLoS Medicine* 3, no. 6 (2006): e195.

New Tactics in Human Rights. “Video as Evidence: To Be Evidence, What Does Video Need?” 2014. https://www.newtactics.org/comment/7427#comment-7427 (accessed February 15, 2017).

Orcutt, Mike. “Are Face Recognition Systems Accurate? Depends on Your Race.” *MIT Technology Review*, July 6, 2016. https://www.technologyreview.com/s/601786/are-face-recognition-systems-accurate-depends-on-your-race/ (accessed February 15, 2017).

Phillips, P. Jonathon, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J. O’Toole. “An Other-Race Effect for Face Recognition Algorithms.” *ACM Transactions on Applied Perception* 8, no. 2 (2011): 1–11.

Piracés, Enrique. “The Future of Human Rights Technology: A Practitioner’s View.” In *New Technologies for Human Rights Law and Practice*, edited by Molly K. Land and Jay D. Aronson, 289–308. Cambridge: Cambridge University Press, 2018.

Porter, Elizabeth G. “Taking Images Seriously.” *Columbia Law Review* 114 (2014): 1687–1782.

Price, Megan, Anita Gohdes, and Patrick Ball. “Documents of War: Understanding the Syrian Conflict.” *Significance* April (2015): 14–19.

Ristovska, Sandra. “The Rise of Eyewitness Video and its Implications for Human Rights: Conceptual and Methodological Approaches.” *Journal of Human Rights* 15, no. 3 (2016): 347–60.

Sasseen, Jane. “The Video Revolution: A Report to the Center for International Media Assistance.” October 28, 2012. http://www.centroperiodismodigital.org/sitio/sites/default/files/Video%20revolution%20FINAL_0.pdf (accessed December 20, 2017).

Shapin, Steven. "Pump and Circumstance: Robert Boyle's Literary Technology." *Social Studies of Science* 14 (1984): 481–520.

Sherwin, Richard K. *Visualizing Law in the Age of the Digital Baroque: Arabesques and Entanglements*. London: Routledge, 2011.

Szeliski, Richard. *Computer Vision: Algorithms and Applications*. New York: Springer, 2011.

Tong, Wei, Yi Yang, Lu Jiang, Shoou-I Yu, ZhenZhong Lan, Zhigang Ma, Waito Sze, Ehsan Younessian, and Alexander G. Hauptmann. "E-LAMP: Integration of Innovative Ideas for Multimedia Event Detection." *Machine Vision and Applications* 25, no. 1 (2014): 5–15.

Wardle, Claire, Sam Dubberley, and Pete Brown. *Amateur Footage: A Global Study of User-Generated Content in TV and Online News Output*. New York: Tow Center of the Columbia Journalism School, 2014.

Weizman, Eyal. *Forensic Architecture: Violence at the Threshold of Detectability*. New York: Zone Books, 2017.

Wexler, Rebecca. "Censorship Through Forensics: Video Evidence in Post-War Crises." In *Global Censorship: Shifting Modes, Persisting Paradigms*, edited by Pranesh Prakash, Nagla Rizk, and Carlos Affonso Souza, 85–108. New Haven, CT: Yale Law School Information Society Project, 2015a.

——. "Convicted by Code." *Future Tense*, October 6, 2015b. http://www.slate.com/blogs/future_tense/2015/10/06/defendants_should_be_able_to_inspect_software_code_used_in_forensics.html (accessed February 15, 2017).