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

Jay D. Aronson

C
. , ,
.
.
,
. A ,
. B ,
, , .

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

emotionally challenging to repeatedly watch videos that depict horrific events like beatings, shootings, torture, suicide bombings, missile attacks, or extrajudicial killings

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 sin-
mis'n4 9e oP2w8w9.6(aais)8t t-375(pa)-3(ino8mCk48(l)-356]hile)1eks-ncidenf 3tlo3us0.gaze. Even

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

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

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 con-

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

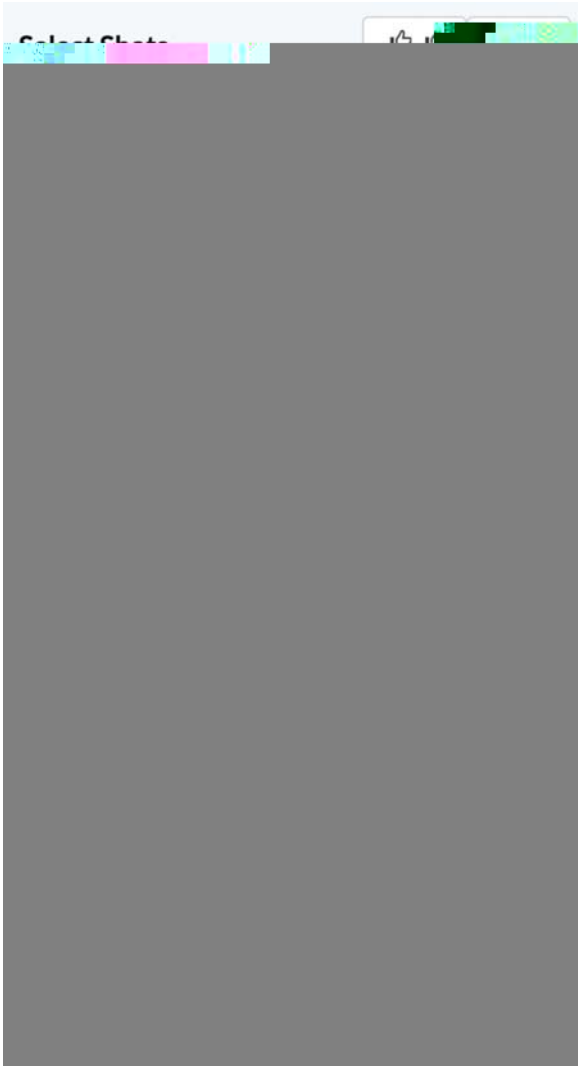

FIGURE 1.
Initial Mortar Launcher Keyframes from the First Video We Found that Contained this Weapon System
Notes: Eighteen total keyframes were taken by E-LAMP and we selected a few of these to build the model for our classifier. We also selected keyframes from other videos to broaden the variety of 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.

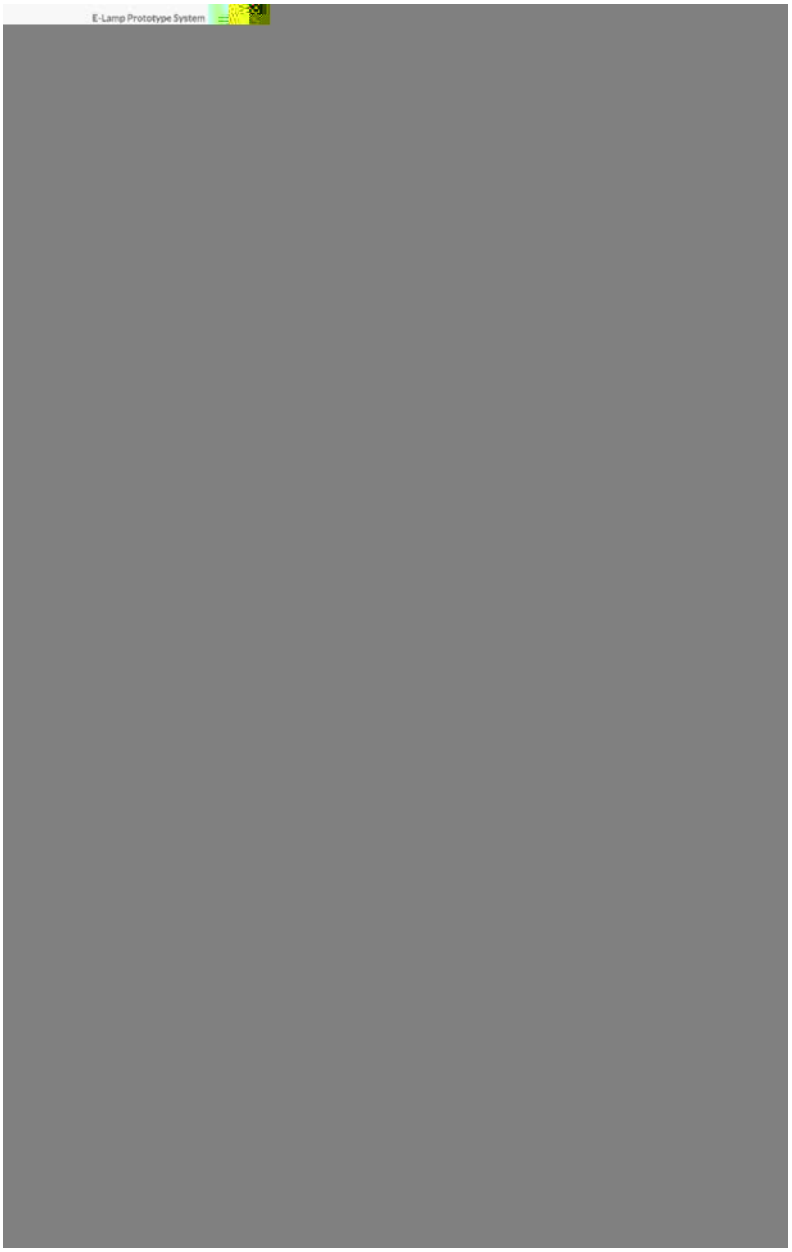

FIGURE 2.
Top 100 Mortar Launcher Videos Out of 476 total. [Color figure can be viewed at wileyonlinelibrary.com]

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

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 investiga-

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 col-

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

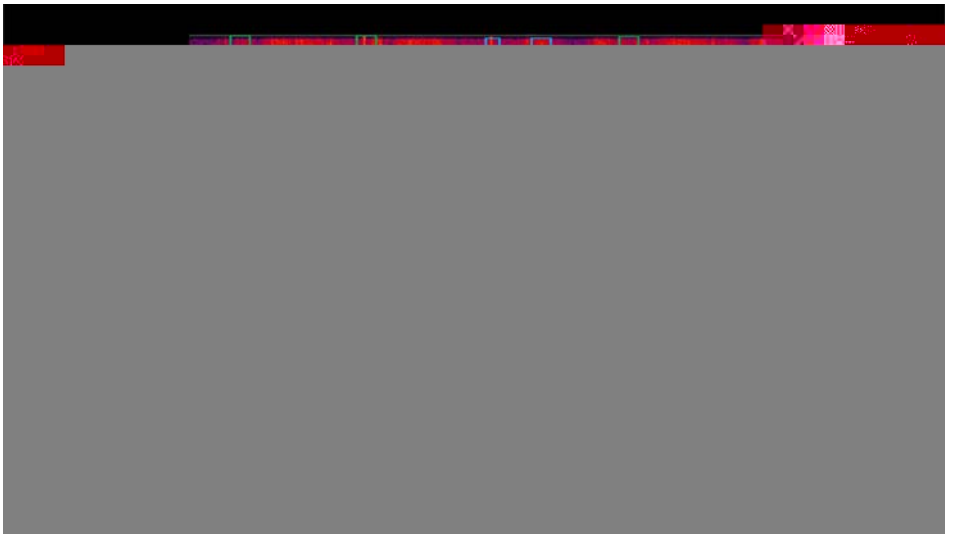

FIGURE 3.
Audio-Based Synchronization [Correction added on XX March 2018 after first online publication: the figure caption has been corrected from "Audio-Based Synchronization Pipeline" to "Audio-Based Synchronization".]. [Color figure can be viewed at wileyonlinelibrary.com]

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.

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

Feigenson, Neal, and Christina Spiesel. *D : D J* 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.situ-research.com/white_phosphorus/index.html.

Garvie, Clare, and Jonathan Frankle. "Facial-Recognition Software Might Have a Racial Bias Problem." *A*

Shapin, Steven. “Pump and Circumstance: Robert Boyle’s Literary Technology.”