Video Analytics for Conflict Monitoring and Human Rights Documentation

Jay Aronson, Shicheng Xu, and Alex Hauptmann
Carnegie Mellon University
July 17, 2015
aronson@andrew.cmu.edu

Abstract: In this article, we describe how a powerful machine learning and computer vision-based video analysis system called Event Labeling through Analytic Media Processing (E-LAMP) can be used to monitor conflicts and human rights abuse situations. E-LAMP searches through large volumes of video for objects (e.g., weapons, military vehicles, buildings, etc.), actions (e.g., explosions, tank movement, gunfire, structures collapsing, etc.), written text (assuming it can be processed by optical character recognition systems), speech acts, and human behaviors (running, crowd formation, crying, screaming, etc.) without recourse to metadata. It can also identify particular classes of people such as soldiers, children, or corpses. We first describe the history of E-LAMP and explain how it works. We then provide an introduction to building novel classifiers (search models) for use in conflict monitoring and human rights documentation. Finally, we offer preliminary accuracy data on four test classifiers we built in the context of the Syria conflict (helicopter, tank, corpse, and gunshots), and highlight the limitations that E-LAMP currently possesses. Moving forward, we will be working with several conflict monitoring and human rights organizations to help them identify the benefits and challenges of implementing E-LAMP into their workflows.

Keywords: Computer Vision, Machine Learning, Human Rights, Multimedia Event Detection, Multimedia Content Analysis

Acknowledgement: The authors would like to thank the MacArthur Foundation and Humanity United for funding this research.

Introduction

In the era of social media, nearly ubiquitous mobile phone coverage, and the spread of Internet connectivity around the world, user generated content is becoming an increasingly important dimension of conflict monitoring and the documentation of war crimes and human rights abuse. As the NGO Witness stated in its seminal 2011 report *Cameras Everywhere*, "Video has a key role to play, not just in exposing and providing evidence of human rights abuses, but across the spectrum of transparency, accountability and good governance. Video and other communication technologies present new opportunities for freedom of expression and information, but also pose significant new vulnerabilities. As more people understand the power of video, including human rights violators, the more the safety and security of those filming and of those being filmed will need to be considered at each stage of video production and distribution. Access to information, technology, skills and networks shapes who can participate–and survive–in this emerging ecosystem of free expression." [1, pg. 16]

Since the publication of *Cameras Everywhere*, there has been an explosion in the availability of documentation of human rights abuses around the world. Journalists, human rights organizations, international institutions, governments, and ordinary people are increasingly finding themselves overwhelmed with massive amounts of visual evidence of suffering and wrong-doing. They must not only determine the veracity of this content, but also how to acquire, archive, analyze, and share this information in a way that is both effective and protective of the rights of those individuals who are present in videos and photographs.

Further, most information extraction from conflict and human rights-related video has been accomplished manually. A researcher will view each relevant video individually, noting whatever particular attributes are of interest. This data will either be expressed in a prose summary or as entries in a database. Such analysis is incredibly time consuming and very expense if people have to be paid to do the work. It is also emotionally challenging to watch numerous videos and extract information from them if the data being gathered deals with issues such as torture, rape, or extrajudicial killings. [2] Additionally, language skills can limit the number of researchers who are capable of carrying out such work.

While this process is acceptable when only a few relevant videos need to be analyzed in a given day, the dissemination of conflict and human rights-related video has vastly outpaced the ability of researchers to keep up with it—particularly when immediate political action or rapid humanitarian response is required. At some point (which will vary from organization to organization), time and resources limitations will necessitate an end to the collection, archiving, and analysis of user generated content unless the process can be automated. This could prevent human rights researchers from uncovering widely dispersed events taking place over long periods of time or large geographic areas that amount to systematic human rights violations.

The Center for Human Rights Science at Carnegie Mellon University is facilitating collaboration among computer scientists, human rights practitioners and social scientists to address these challenges. The long-term goal of this partnership is the creation of "Human Rights Media Central," a data management system that will enable the discovery, acquisition, archiving, authentication, organization, analysis (including the search for duplicates or near duplicates), utilization, and sharing of human rights-related user generated content including images, video recordings, and text. Many individuals and organizations are working on various aspects of this challenging problem, and time seems right to harmonize these efforts under a single roof—or at least ensure that the products created are interoperable, adhere to generally accepted principles and standards (many of which remain to be agreed upon), and produce outputs that can be integrated into the workflows of the organizations and advocates who use such media in their everyday activities. Ideally the system (or systems) can be tailored to the needs of the user, but still provide the kind of safety/security, chain of custody assurance, and analytic capabilities that all users need and deserve.

The Event Labeling through Analytic Media Processing (E-LAMP) System

As the first step of this ambitious agenda, we are in the process of providing a test group of our human rights partners with access to a powerful machine learning and computer vision-based video analysis system developed at Carnegie Mellon University by Alex Hauptman and colleagues called Event Labeling through Analytic Media Processing (or E-LAMP). [3] This work on video analysis began in 1994 as the Informedia Digital Video Library project, with the mission to make video as easily accessible and searchable as text. Informedia originally dealt with educational video materials such as documentaries to support educational goals, and soon expanded to the analysis of broadcast news. Early capabilities included news story detection, topic identification, and labeling of key persons, organizations and locations (named entities) as well as video summarization. From there, offshoots of the Informedia project worked on analysis of user video from body cameras, and surveillance monitoring in nursing home environments (CareMedia). Throughout this evolution, the research integrated Spanish, Serbian, Chinese and Arabic speech recognition; video Optical Character Recognition for certain alphabets, as well as provided translations via Google Translate. [4]

Serendipitously for the human rights community, the latest iteration of E-LAMP has been adapted to work on videos uploaded to social media sites or shared via electronic communications. E-LAMP does not require any metadata to function, and is capable of quickly searching through thousands of hours of video for relevant visual objects, semantic concepts, speech, or text. [5] It also includes an interface to allow the user to interactively build classifiers for the objects, scenes or activities in which they are interested.

As we demonstrate below, E-LAMP has the potential to aid human rights researchers in their monitoring and documentation efforts. In particular, the system can be

set up to search for particular objects (e.g., weapons, military vehicles, buildings, etc.), actions (e.g., explosions, tank movement, gunfire, structures collapsing, etc.), locations, written text (assuming it can be recognized by optical character recognition software), speech acts (either words or patterns of speech), human behaviors (running, crowd formation, crying, screaming, etc.), and identification of particular classes of people such as soldiers, children, or corpses. It also seems likely that identification of individuals will eventually become possible as facial recognition continues to improve. There are many possible future uses of search capacity like this as the technology continues to improve. In the short term, it can provide a triage system to allow overburdened researchers to focus their efforts on videos that have a high likelihood to contain information of relevance to them, or allow them to screen through large corpuses of video for the information they need. Videos not flagged can be screened during times of lower time pressure, or by staff with less specialized knowledge of a conflict or situation. In the longer term, it will also assist human rights researchers in their documentation of systematic abuses and war crimes that are geographically and temporally diffuse.

Two Notes of Caution

Before moving forward, two notes of caution are in order. First, no video analysis system can replace the expert judgment of a knowledgeable human being when it comes to understanding what is happening in a video. Used properly, a technology like E-LAMP can cut down the amount of time it takes for human researchers to find the content that is relevant to an investigation so that they can concentrate on the challenging work of interpretation that only humans can currently do. Such interpretation rests squarely on topical and regional expertise, as well as corroboration by eyewitnesses and others with direct knowledge of an incident or situation. It would be foolish to suggest that such systems can replace human researchers—their current role is to increase the efficiency of conflict analysis and human rights fact-finding and, hopefully cut down on the emotional toll that watching endless loops of suffering, death, and cruelty takes on those who dedicate themselves to ending human rights abuse and impunity. Second, in almost all cases, user generated content is the starting point for an investigation, not an end point. No human rights researcher should ever take any data source, particularly one as complex as social media content, at face value as "the Truth." Rather, the researcher must take the opposite approach: assume that each individual piece of content is potentially false or misleading and that the collected dataset as a whole is biased and incomplete. [6,7] Social media content must be authenticated, subjected to rigorous investigative scrutiny, and placed in the proper historical, political, and cultural contexts. Ideally it should also be triangulated with as many other sources of information as possible.

How E-LAMP Works

The principal function of E-LAMP is to autonomously detect a predefined event in a large collection of video material of varying quality. For our purposes, an event is defined as "a complex activity occurring at a specific place and time which involves people interacting with other people and/or object(s)." [2, pg. 1] Such 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—hence early work in this field tended to focus on video from stable and predictable environments, such as fixed camera surveillance video or sporting events bound by clear rules, environments (the basketball court, the soccer field, the tennis court, etc.) and a limited number of actions (e.g., dribble, pass, shoot, hit, etc.). In these situations, particularly in the context of surveillance video, a limited set of actions can be fully modeled and classifiers can be developed to detect both these actions and anomalous ones. [8]

Social media video presents a much richer challenge on a variety of levels: to begin with, the camera is unconstrained in time and space and generally offers something akin to first person vision. The videographer (often an amateur) is in motion and reacting to the events he or she is capturing on camera. Second, social media contains an almost infinite number of situations (even if some, such as silly cat behavior or how-to videos tend to appear more frequently than others) that are not bound by clear-cut rules or spatial environments. Finally, YouTube captures an almost infinite variety of human and non-human behavior—in order to effectively mine this resource, an analyst needs to be able to develop novel classifiers on the fly.

The E-LAMP system directly addresses these challenges. 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 and a set of null videos that depict other unrelated activities. E-LAMP analyses these videos for a variety of different features (most of which are totally unrelated to the way that human would carry out the same process), which can be combined into a computational machine learning model of the relevant action or event, then delves into a larger collection of videos to look for other potential examples of this model. It then 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% accurate for a variety of reasons), this set of patterns is labeled as a classifier, or event kit, for the particular action. As defined by the U.S. National Institute for Standards and Technology, a classifier includes the following characteristics:

- An event name which is an mnemonic title for the event
- An event definition which is a textual definition of the event
- An evidential description is a textual listing of attributes that are indicative of an event instance. The evidential description provides a notion of some potential types

of visual and acoustic evidence indicating the event's existence but it is not an exhaustive list nor is it to be interpreted as required evidence

- A set of illustrative video examples each containing an instance of the event. The examples are illustrative in the sense that they help to form the definition of the event but they do not demonstrate all possible variability or potential realizations. [quoted in 9]

This classifier, which can be visual, aural, or semantic, 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 scraped footage. It also makes it possible to gain multiple perspectives on a single event.

In addition to visual classifiers, E-LAMP currently has the ability to perform speech recognition in English and Arabic (together with preliminary versions of Spanish, German, French, Mandarin Chinese, Cantonese, Turkish, Pashto, 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 and rendered searchable. Speech is also analyzed by automated speech recognition software and is similarly rendered searchable. Our preliminary testing in the context of human rights related videos has shown that the accuracy of OCR varies tremendously based on the quality of the video, the quality of the actual text, and the alphabet being used, while the accuracy of speech recognition varies depending upon sound quality, the dialect or accent of the speaker, and the state of speech recognition systems for the language being spoken.

In essence, E-LAMP functions by literally watching the video in its entirely one time through, dividing it up into various scenes (i.e., anytime there is a significant shift in scene or action), and then taking a visual snapshot (a "keyframe") at the midpoint of each section. The feature detection described above is then performed on these frames so that they can be compared with classifiers for the targeted event.

The most computationally intense aspect of this system is the initial processing of the videos. Based on the current performance of the system, 800 hours of raw video (which is the approximate size of our current conflict test collection) takes approximately 1243 hours to process. At least some of this processing can be done in parallel on Graphical Processing Units (GPUs), so the actual time required to process at batch of videos this size is significantly shorter—taking a week or two depending on available computing resources and technical expertise (many of the steps must currently be done manually by an individual with a reasonable level of computer training). At the current cost of $2.34/hour of video, processing 800 hours of video would cost approximately $1,872 just for computing resources. Further, approximately 180-200 GB of storage (either cloud or hard

drive-based) is required for a collection of this size. Over time, we hope to significantly shorten the amount of time needed for processing and also automate the process as much as possible.

Building a Classifier for Mortar Launchers

In order to provide a first test of the capacity of system to aid in human rights investigations, we took a subset of around 500 videos that focused on events taking place in Aleppo, Syria in late 2013. In this example, we imagined that a researcher wants to understand how well-armed particular rebel groups were within the region, but did not have access to on-the-ground reports of the availability of weapons in the area. In this case, she could search available conflict videos for a variety of weapons such as machine guns, surface-to-air missiles, tanks, etc. Given the nature of the conflict in Syria, rebel groups routinely film their military exploits and regularly report about their caches of weapons. It is of course important to recognize that all such videos are public relations ploys, and that no group makes public 100% of their actions (especially those that highlight their weaknesses or show them being defeated). Further, the absence of a phenomenon in a video collection does not imply or even guarantee absence of that phenomenon on the ground, and one cannot make statistical calculations of any sort based on available social media reports because they are only a convenience sample of data. [7] 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 a particular weapon system in a particular region.

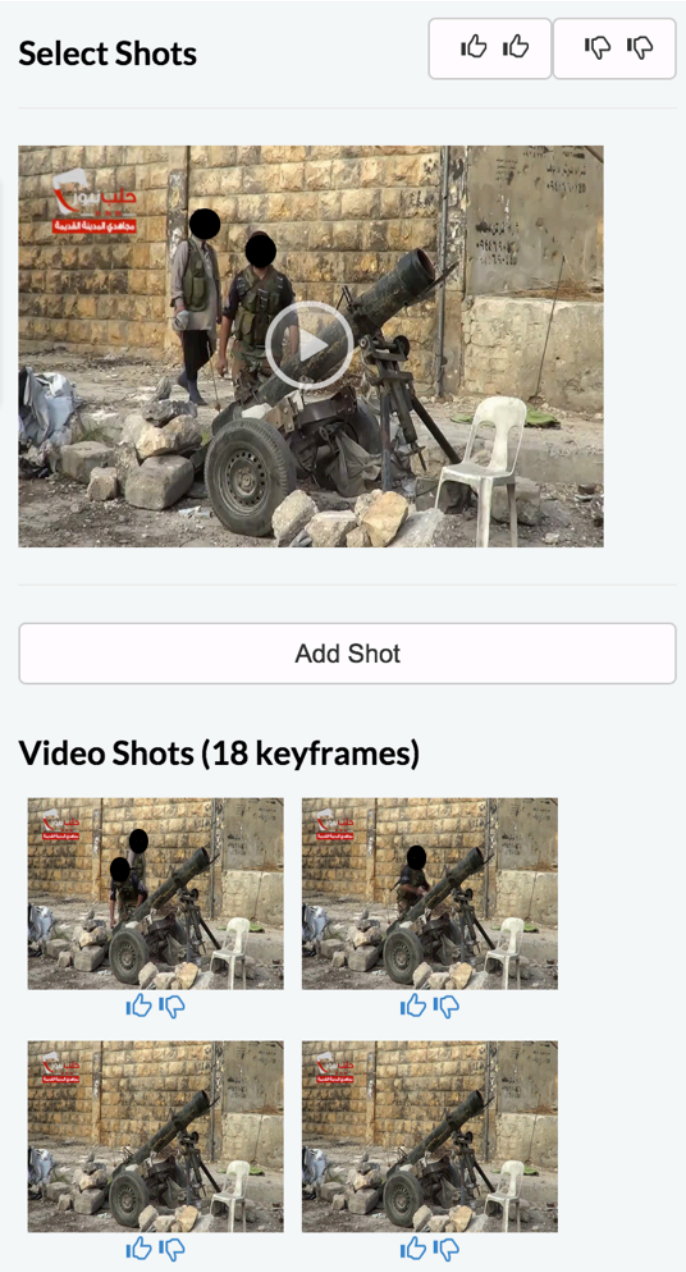
At the moment, there are over 2500 pre-existing semantic concepts available for use in searching a video collection. Because the system was not trained for conflict and human rights analysis, though, classifiers for much of the material of interest to researchers in this domain will have to be built from scratch. Thus, when we first uploaded the Aleppo video collection, there was no semantic concept for "mortar launcher," or most of the other weapons system mentioned above. 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 Syria 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, Optical Character Recognition (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. (This process assumes that the system can understand the dialect the words are spoken in or the script they are written in. Because this is not always a safe assumption, particularly in the context

of Arabic, we are actively working on enhancing the language and text processing capacities of the system.) The researcher can also manually select a few videos of mortal launchers that she knows are in the larger video collection, which is how we went about the process.

Assuming that investigator has a few videos containing mortar launchers, she then can either play the videos in their entirety and select particular "shots" of interest (i.e., moments when the mortar launcher is visible or heard being launched), or use "keyframes" (the thumbnails automatically extracted by the system for video analysis purposes with motion and audio information embedded within them) 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 extracted from these videos shots to create a classifier that will become the basis for a new semantic concept of "mortar launcher."

Figure 1: Initial mortar launcher keyframes from the first video we found that contained this weapon system. 18 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 didn't overly limit the machine's final model.



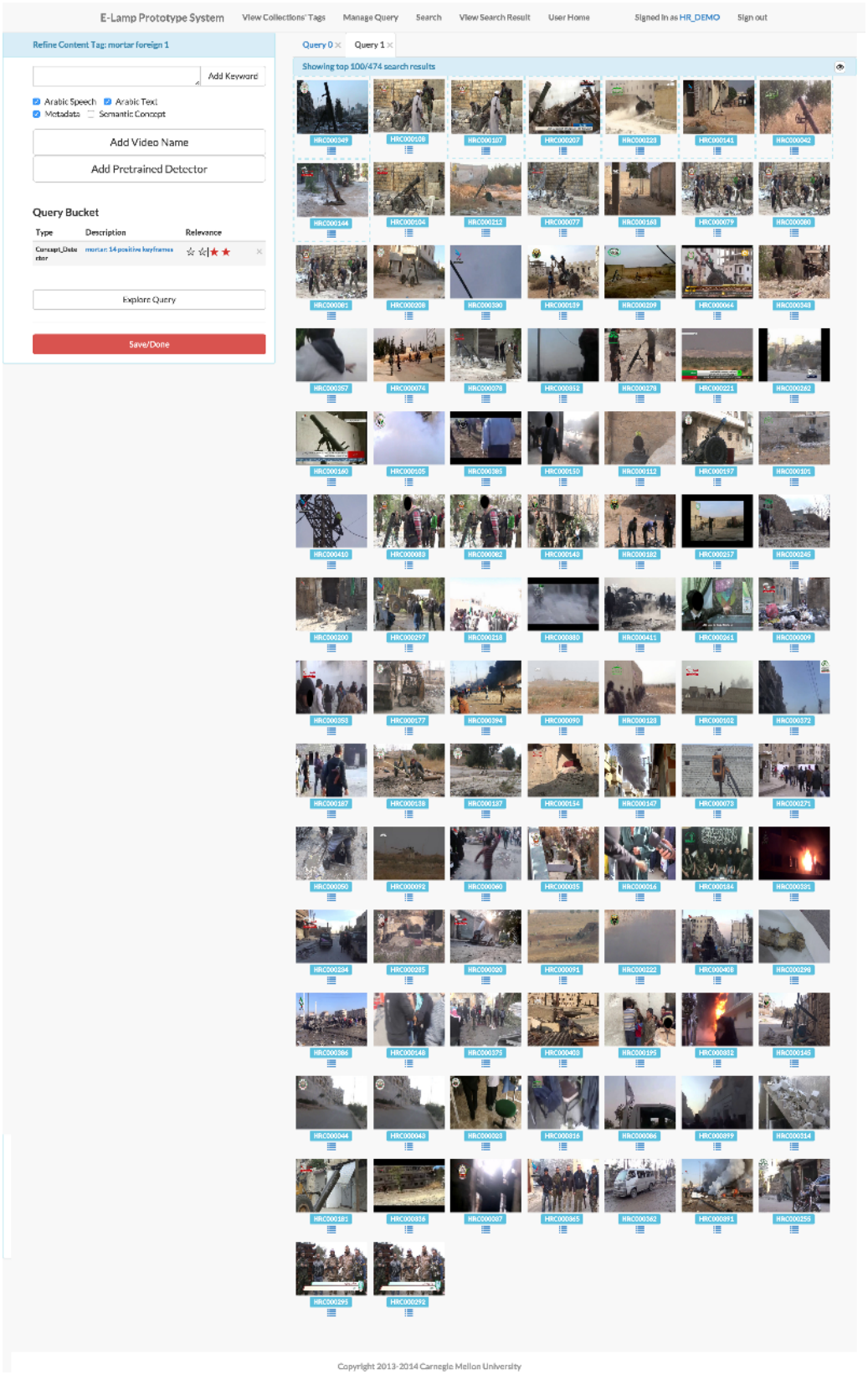
The operator then runs a preliminary test of the newly trained classifier to determine how well it works. She can refine the classifier in two ways. After the training process is finished, the training panel will be expanded to visualize the semantic concept model (Figure 2). The video shots are arranged so that the examples that may confuse the system are arranged near the decision boundary. The user can examine these videos to see if there is are false positives or false negatives, then correct the labels and retrain the

model. The users can also simply apply this model to the video collection and provide additional feedback through the tagging of more positive or negative video shots.

Figure 2: "Mortar Launcher" Semantic Concept Model

It will likely take a few iterations to minimize examples that are incorrectly detected by the model (particularly those that look or sound similar to the phenomenon in question). In the case of the mortar launcher example, the model misidentified things like power lines, truck mounted anti-aircraft guns, and other linear objects with similar backgrounds as the 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 a lot of positive but lower confidence cases. A less sensitive classifier will capture a greater percentage of positive videos cases (both high and low confidence), but will also capture a lot of 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 (Figure 3).

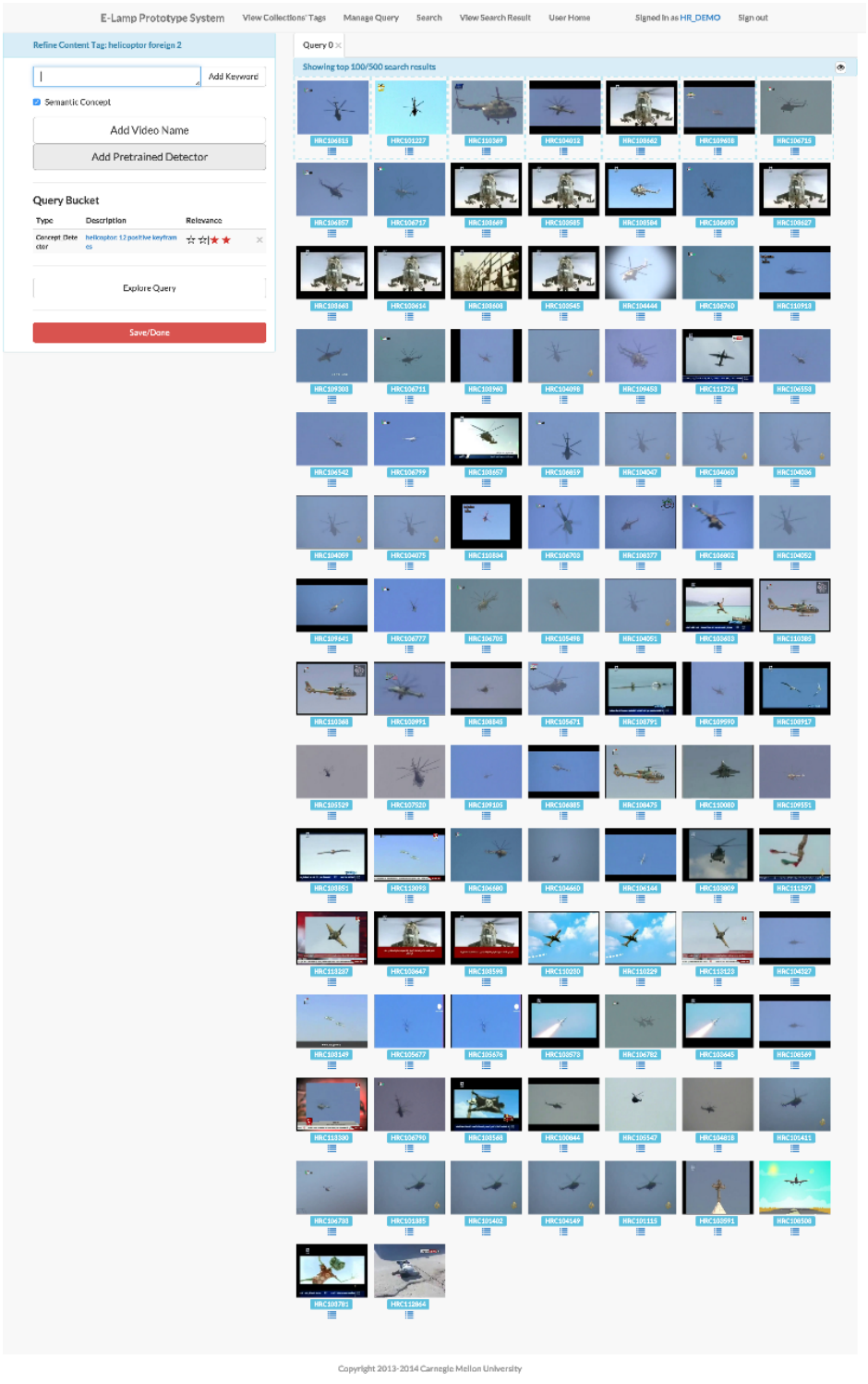
Figure 3: Top 100 Mortar Launcher Videos out of 476 total (presence of many non-mortar launcher videos is because of small sample size in initial test run)


E-Lamp Prototype System
View Collections' Tags
Manage Query
Search
View Search Result
User Home
Signed In as HR_DEMO
Sign out
Refine Content Tag: mortar foreign 1
Add Keyword
Arabic Speech
Arabic Text
Metadata
Semantic Concept
Add Video Name
Add Pretrained Detector
Query Bucket
Type
Description
Relevance
Explore Query
Save/Done
Query 0
Query 1
Showing top 100/474 search results
Copyright 2013-2014 Carnegie Mellon University

Three Test Cases

Once we were confident that E-LAMP would work on videos from the Syria conflict, we decided to move to a larger collection of 13,570 videos and try to find more complicated phenomena, including helicopters, corpses, and gunshots. These videos were a subset of a much larger collection held by the Syria Justice and Accountability Center, one of our human rights partners. They were all retrieved from social media between 2012 to mid 2014 from the channels of a variety of media outlets, activists, and armed groups. Building the helicopter classifier was relatively straightforward and followed the same process described above. This classifier is useful in identifying barrel bomb videos, among others. We achieved excellent results after three rounds of training (Figure 4). The most common misidentification 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 an gecko mascot that appears to be falling from the sky with its four limbs spread out in an American insurance company advertisement) that show up toward the bottom of the top 100 results.

Figure 4: Helicopter Top 100 Search Results (see next page). Note the presence of several airplanes as well as the large number of duplicate images. The duplication suggests either that several copies of the same video are in our test collection or that individuals and groups posting to social media are reusing the same footage.


E-Lamp Prototype System
View Collections' Tags
Manage Query
Search
View Search Result
User Home
Signed In as HR_DEMO
Sign out
Refine Content Tag: helicoptor foreign 2
Add Keyword
Semantic Concept
Add Video Name
Add Pretrained Detector
Query Bucket
Type
Description
Relevance
Explore Query
Save/Done
Query 0
Showing top 100/500 search results
Copyright 2013-2014 Carnegie Mellon University

The "corpse" detector was similarly successful. To build this classifier we selected several images of bloodied bodies with visible faces in a horizontal pose with no movement, closed eyes, and an expressionless face. On the first iteration, only 5 out of the top 100 hits were incorrect. 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 a couple of images of what appear to be pink blossoms against dark green leaves (search results available upon request). We believe that we could have refined the model further and prevented these mistakes from occurring, but did not do so because we wanted to limit our exposure to images of death. It is important to note that we likely missed many cases of fully covered corpses (i.e., with unexposed faces), so a separate classifier would have to be built for those cases. We also did not check our results in detail to determine whether any of the videos actually depicted gravely wounded, but not deceased, individuals.

The gunshot detector also worked very well, but took five iterations to achieve an acceptable level of accuracy because of the number of sounds that mimic gunshots in our test video collection. We built this classifier the same way as the others, but used sound features rather than visual features and searched only for sounds. We did not use any visual information to build the classifier. Of the top 100 search results, 66 were correct. The model detected a variety of tat-tat-tat type sounds as well as a microphone pop in an armed formation video that sounded like a single gunshot.

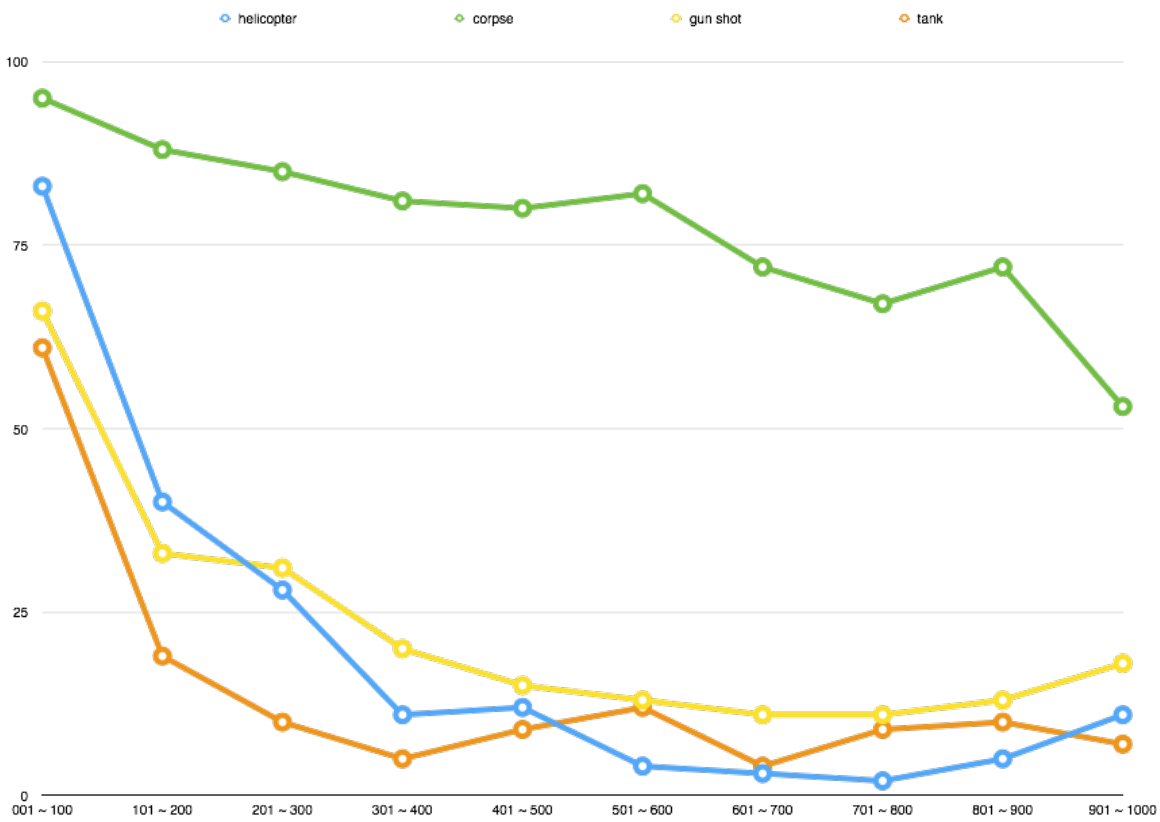
Accuracy

It is impossible to make a blanket statement about the accuracy of E-LAMP in the human rights context because the success of each classifier relies heavily on how it was trained, the desired stringency of the search, and the quality of videos in a given collection. That said, we still think it is important to provide readers with a general sense of how the system worked on the 13,570 videos that we have from the Syria conflict. It should be noted that our quantitative analysis is rather subjective (in that it depends upon individually checking each video returned after a search to determine whether or not it is actually the thing being sought after) and is not statistically rigorous. We hope to undertake more scientific studies of the accuracy of the system, as well as analyses of how to set stringency levels to achieve the best results in a given search, in the future. Further, our preliminary analysis says nothing about how the system will work for more complex phenomena (such as one human or a group of humans beating another human or group of humans with a blunt object; remnants of a barrel bomb or other form of munitions; or a particular sound less distinct than a gunshot) or on other video collections.

For now, what we can say is that the E-LAMP system did a good job of detecting the phenomena for which we built classifiers in the video collection we used (see Figure 5). In our experience with the case studies described here, and several other classifiers not discussed in this paper, the top 50 results tend to be very accurate with sufficient training,

while with the next 50 tend to vary in accuracy depending on the overall number of positive videos, the classifier and the quality of the video collection. Once high confidence results have been detected, the number of positives detected drops quickly for the classifiers we tested, but not to zero. In other words, there are still a small, but potentially significant, number of obviously positive results that will be ranked very low by the machine learning model. The reasons for such false negatives probably vary based on the individual characteristics of the video and we do not attempt to assign any causation at this point. One important conclusion that can be reached from this preliminary analysis is that E-LAMP should not be used alone to undertake statistical analyses of trends or patterns in datasets, not only because the datasets being used are likely biased and incomplete, but also because E-LAMP will almost always misclassify a percentage of videos at a rate that will likely differ in each case. The use of E-LAMP for trend and data analysis will have to wait for formal validation of the system, the development of a mechanism to determine error rates within a particular collection, and a better understanding of how to train classifiers and set thresholds for calling positives.

Figure 5: Number of Positive Results per 100 total results for first 1000 videos returned by the machine learning model for corpse, tank, helicopter, and gunshots in a collection of 13570 videos:



Next Step: Pilot Implementation

In order to test the utility of E-LAMP, we are partnering with several human rights organizations to help them implement it into their workflows and to determine the extent to which it actually improves their ability to defend human rights. These test cases will then help us refine the technology and determine how best to roll it out the larger community. We will be focusing on four major challenges:

1) *Detection of objects, text, and sound* in a large corpus of video material for human rights-related concerns. In the context of Syria, for example, we have a large historical database of videos (around 300,000) that have been scraped from various social media sources. These videos depict a wide variety of circumstances related to the conflict, and contain a wealth of information for researchers to mine, including:

- Protests
- Armed group formation videos (number of fighters, weapons, requests for assistance, membership drives)
- Active conflict (bombardment, fighting, shooting, bombing from multiple perspectives [perpetrator, target, witness])

- Human toll (hospital scenes, injuries, casualties, evidence of torture, direct depiction of torture, family members, refugees, etc.)
- Destroyed infrastructure (buildings, roads, and other facilities)
- Pleas for help
- Humanitarian organization videos
- Interviews with leaders
- Government information and propaganda

We will work with human rights groups focused on the Syria conflict to help develop visual and sound classifiers related to these types of objects, events, and actions.

2) *Video triage*. Another major problem faced by human rights researchers is that during times of intense conflict, political instability, or disaster, it is difficult to keep up with the sheer volume of videos that are produced in a given region. Working with several human rights monitoring and advocacy organizations, we are developing a system that will semi-autonomously analyze new content on social media platforms and flag potentially relevant material for review. The capacity of E-LAMP to identify a video's typology (i.e., which of the above scenarios are displayed in a particular video) using an "event query" (not described in this paper) will be useful in this context. The most relevant videos can then be automatically sent to waiting researchers for rapid human analysis and action. It should be noted that this process is currently dependent upon pre-existing search criteria (whether by keyword, user, or regional designation). Our system cannot autonomously analyze ALL video material uploaded to social media sites.

3) *Workflow Integration*. Technology and data are both meaningless unless a researcher can actually use them and make their efforts more efficient and effective. As such, one of the goals of this project is to help understand the ways that video analytics can be integrated into the workflows of a variety of types of human rights organization. We will work with our human rights partners to experiment with various ways to integrate E-LAMP's video analytics capacity into their daily activities.

4) *Practical Issues: cost, processing efficiency, and user interface*. Over the long term, we will also work to bring the costs associated with E-LAMP down, decrease the time needed to process the video, and generally automate the system and simplify the user interface as much as possible to make it more user friendly for organizations that do not have an dedicated IT personnel or computing expertise.

Monitoring and Evaluation

The technology described here has numerous technical limitations, which have been clearly articulated throughout the article. For those organizations that can make use of the

system in the near term, their staff must expect that there will likely be a significant rate of false positives and false negatives for complex classifiers. Researchers must develop plans to ensure that this does not lead to large amounts of wasted time or large numbers of relevant videos that get lost in the shuffle. Further, a system like this one is no substitute for human analysis and judgment. In reality, at best E_LAMP gives human researchers more time to analyze content and make determinations about what ought to be done in response to events captured on film or what information can be extracted from a video for use in justice and accountability efforts. There is always a risk that a system like E-LAMP won't increase the efficiency or effectiveness of a researcher or organization, but will bog them down in technical trouble-shooting or videos that do not provide them with relevant information. Finally, the costs, timeframe, and requirements of computing expertise are still too high for many if not most human rights organizations to use E-LAMP on a daily basis.

As such, as implementation and integration of E-LAMP progresses we will be monitoring and evaluating the effectiveness of this technology along the following lines:

1) Does E-LAMP increase the amount of information human rights organizations are able to mine from video materials?
2) If not, is the problem caused by lack of information extraction, too much information, too much error/noise, or some other issue?
3) If so, does this information lead to tangible gains in the organization's ability to engage in advocacy and accountability processes?
4) Do the outputs of E-Lamp integrate into the existing workflows of the organization?
5) If not, what can be done to improve this process?
6) Does E-LAMP make the work of human rights organizations more efficient—both in terms of financial resources and human capacity? (how to measure??)
7) Along these lines, does E-LAMP free up human personnel to do more high-level analysis or advocacy, or does it create an added burden on personnel that takes away from these activities?

Conclusion

While the overall accuracy of E-LAMP varies depending on what is being search for and the quality of the video collection, our initial tests suggest that the system can be a powerful tool for use by the human rights community when its limitation are sufficiently understood and taken into account. E-LAMP seems poised to help improve the efficiency of human rights documentation efforts by speeding up the analysis of large amounts of video, and aiding in the development of video triage systems, but it cannot replace human researchers or make policy decisions. Researchers implementing this system will need to take time to understand the strengths and limitations of this technology and plan their workflow accordingly. For instance, they will have to determine how much time and resources need to be put into looking for relevant videos that the system may have missed.

Finally, it is important to note that E-LAMP is the first step in a broader effort to give the entire human rights community access to a broad suite of technologies to make it easier to gather, archive, analyze, and utilize relevant video content—it is of limited value when used in isolation.

References

1. S. Padania et al., *Cameras Everywhere: Current Challenges And Opportunities At The Intersection Of Human Rights, Video And Technology* (WITNESS, 2011), http://www.ohchr.org/Documents/Issues/Opinion/Communications/Witness_1.pdf.
2. EyeWitness Media, "Making Secondary Trauma a Primary Issue," June 15, 2015, https://medium.com/@emhub/making-secondary-trauma-a-primary-issue-24a802abe461.
3. W. Tong et al., "E-LAMP: integration of innovative ideas for multimedia event detection," Machine Vision and Applications, vol. 25, no. 1, 2014, pp. 5-15, doi: 10.1007/s00138-013-0529-6.
4. "Informedia: Digital Video Understanding," Carnegie Mellon University, 1994-2010, http://www.informedia.cs.cmu.edu.
5. S. Xu, et al., "Incremental multimodal query construction for video search," ACM International Conference on Multimedia Retrieval, 2015.
6. C. Silverman (ed.), *Verification Handbook for Investigative Journalism* (European Journalism Centre, 2015), http://verificationhandbook.com/downloads/verification.handbook.2.pdf.
7. M. Price, A. Gohdes, and P. Ball, "Documents of War: Understanding the Syrian Conflict," Significance, April 2015, pp. 14-19.
8. A. Adam, et al., "Robust real-time unusual event detection using multiple fixed-location monitors," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 3, 2008, pp. 555–560, doi: DOI: 10.1109/TPAMI.2007.70825.
9. P. Over et al., "Trecvid 2010—an introduction to the goals, tasks, data, evaluation mechanisms, and metrics," National Institute of Standards and Technology, 2010, http://www-nlpir.nist.gov/projects/tvpubs/tv10.papers/tv10overview.pdf