

6-2016

# Video Synchronization and Sound Search for Human Rights Documentation and Conflict Monitoring

Junwei Liang  
*Carnegie Mellon University*

Susanne Burger  
*Carnegie Mellon University, sburger@cs.cmu.edu*

Alex Hauptmann  
*Carnegie Mellon University, alex@cs.cmu.edu*

Jay D. Aronson  
*Carnegie Mellon University, aronson@andrew.cmu.edu*

Follow this and additional works at: [http://repository.cmu.edu/chrs\\_reports](http://repository.cmu.edu/chrs_reports)

---

This Technical Report is brought to you for free and open access by the Center for Human Rights Science at Research Showcase @ CMU. It has been accepted for inclusion in CHRS Reports by an authorized administrator of Research Showcase @ CMU. For more information, please contact [research-showcase@andrew.cmu.edu](mailto:research-showcase@andrew.cmu.edu).

# Video Synchronization and Sound Search for Human Rights Documentation and Conflict Monitoring

CMU Center for Human Rights Science (CHRS) Technical Report

Junwei Liang  
Susanne Burger  
Alex Hauptmann  
Jay D. Aronson

June 2016

**Abstract:** In this technical report, we present a powerful new machine learning-based audio processing system that enables synchronization of audio-rich video and discovery of specific sounds at the frame level within a video. This tool is particularly useful when analyzing large volumes of video obtained from social media and other open source Internet platforms that strip technical metadata during the uploading process. The tool creates a unique sound signature at the frame level for each video in a collection, and synchronizes videos that are recorded at the same time and location. The use of this tool for synchronization ultimately provides a multi-perspectival view of a specific event, enabling efficient event reconstruction and analysis by investigators. The tool can also be used to search for specific sounds within a video collection (such as gunshots). Both of these tasks are labor intensive when carried out manually by human investigators. We demonstrate the utility of this system by analyzing video from Ukraine and Nigeria, two countries currently relevant to the work of Center for Human Rights Science collaborators.

**Keywords:** Audio Signal Processing, Machine Learning, Multi-Camera Video Synchronization, Conflict Monitoring, Human Rights Documentation

**Acknowledgements:** The authors would like to thank the MacArthur Foundation, Oak Foundation, and Humanity United for their generous support of this project.

## **Background on the CHRS Media & Human Rights Program**

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 expensive 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.

To address these challenges, the Center for Human Rights Science is facilitating collaboration among computer scientists, human rights practitioners, and social scientists through our Media & Human Rights Program. The long-term goal of this partnership is to create a set of tools and methods that will enable the discovery, acquisition, archiving, authentication, organization, analysis, utilization, and sharing of human rights-related images and videos. These tools and methods will be as interoperable as possible, and generate outputs that can be integrated into the workflows of the organizations and

advocates who use such media in their everyday activities. Ideally they 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 the human rights community needs and deserves.

### **Overview of Audio Processing System**

In this technical report, we present a powerful new machine learning-based audio processing system that enables both synchronization of multiple audio-rich videos of the same event, and discovery of specific sounds (such as wind, screaming, gunshots, airplane noise, music, and explosions) at the frame level within a video. The tool creates a unique “soundprint” for each video in a collection, synchronizes videos that are recorded at the same time and location based on the pattern of these signatures, and also enables these signatures to be used to locate specific sounds precisely within a video. The use of this tool for synchronization ultimately provides a multi-perspectival view of a specific event, enabling more efficient event reconstruction and analysis by investigators.

Video analysis techniques have traditionally focused on the visual domain. Given that vision is the most data-rich sensor for humans, it makes sense that machines would also be able to extract significant semantic information from images and videos. However, there are certain situations in which visual sensors fail to provide reliable information. For example, when an object is occluded, poorly illuminated, or not visually distinct from the background, it cannot always be detected by computer vision systems. Further, while computer vision can provide investigators with confirmation that a particular video was shot from a particular location based on the similarity of the background physical environment, it is less adept at synchronizing multiple videos over time because it cannot recognize that a video might be capturing the same event from different angles or distances. In both cases, audio sensors function better so long as the relevant videos include reasonably good audio.

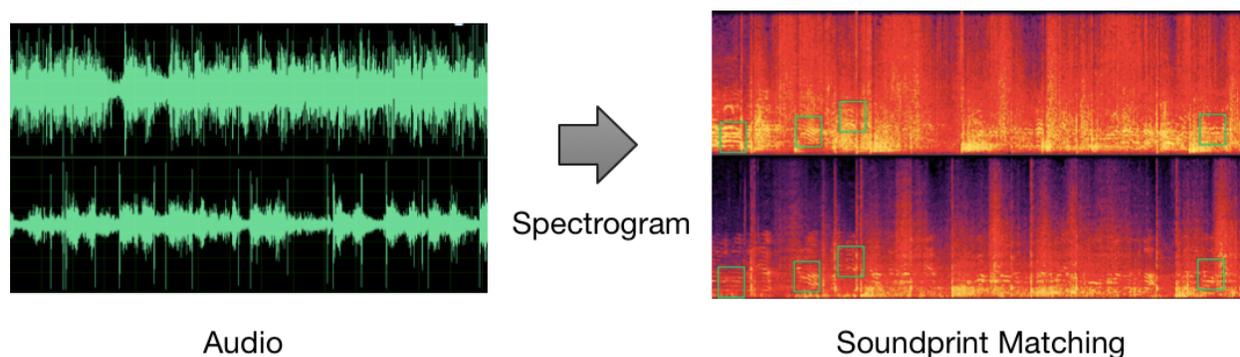
### **Multi-Perspectival Video Synchronization**

After experimenting unsuccessfully with vision-based approaches to synchronization, we decided to test a variety of audio approaches. Our early attempts were hampered by a variety of challenges, including: the high level of background noise present in most videos related to conflict and human rights abuse; the fact that this background noise may differ dramatically depending on where the person filming an event is located (e.g., at ground level; on the roof of a tall building; or in or near an idling vehicle); differences in volume or sound intensity caused by the varying quality of microphones on the most widely carried mobile phones; and the offset in sound caused by differing distances to a source of noise based on where various people are filming from (sound travels much slower than light and such differences can be seen even within relatively short distances). Thus, we determined that any system we created had to be able to take all of these difficulties into account.

In general terms, our audio-based video synchronization system takes a query video, develops a soundprint based on patterns of spectral signatures, and automatically finds other processed videos in a collection that share a similar pattern. More technically, the system first pre-processes all videos by segmenting them into discrete, continuously recorded clips. It does this because many videos in the human rights and conflict-monitoring domains are edited compilations of discontinuous video segments. The system

then analyzes the frequencies present in the audio and produces a spectrogram to represent each clip. Our algorithm recognizes the spectral signatures of a standardized vocabulary of sounds that we predetermine based on the collection we are analyzing (in this case wind, screaming, gunshots, airplane noise, speeches from a loudspeaker, music, and explosions, among others). The algorithm then compares the sequence of these features in each clip to all others and looks for reasonable matches. A sample spectrogram can be seen in Figure 1.

The system also generates a match score to give the human analyst a sense of how reliable the match is. All matches are subject to human review in order to ensure that the system was not fooled by events that may have similar spectral signatures but are actually not the same (e.g., squealing tires and screaming). The system also allows the human analyst to manually adjust the offset of videos that are matched correctly based on audio but do not align perfectly visually. This situation will arise when videos of the same event are shot from different distances - thus the sounds will take different times to get to each recording device.



**Figure 1. Audio Synchronization** - *Audio Soundprint Matching analyzes audio signals and matches similar spectral features (marked here with green squares).*

### **Example: Synchronizing Ukraine Euromaidan Protest Videos**

The Center for Human Rights Science and our collaborator SITU Research recently received a request from Ukrainian human rights practitioners working with families of protesters killed during the 2013-2014 Euromaidan Protests to organize and analyze videos from those events. These protests started over the Ukrainian government's decision to pull back from European integration and seek to strengthen diplomatic ties with Russia, and quickly evolved into a broader movement over democracy, economic security, and corruption. Our partners wanted to locate every video available in their collection of the moments before, during, and just after a specific set of killings. They wanted to extract information from these videos, including visual depictions of these killings, whether the protesters in question were an immediate and direct threat to the security forces, plus any other information that could be used to corroborate or refute other forms of evidence or testimony available for their cases.

Our Ukrainian partners originally hoped to manually synchronize more than 65 hours of video available from the morning of February 20, 2014, when the victims whose families they are representing were killed, and find those videos from the time and place

where they were killed. In order to do so, they worked with an activist who was intimately familiar with the events of Euromaidan and the geography of Kiev. Working full time over the course of several months, she was only able to stitch together a small percentage of the total video using visual and audio cues in the recordings. She compiled this work in a 9-channel video that stretches to 4:37:19 hours total, but contains large gaps throughout. While this work produced an impressive synchronization of a portion of the video content, it became clear that it was prohibitively time consuming to analyze all of the available footage in this manner.

Responding to the request, we developed the audio analysis system described in this technical note. Once we were confident in the robustness of the system, we set to work on the 520 videos we received, which represented a total length of 65 hours. Because many of the videos were edited combinations of many scenes, we segmented these videos into unique, unedited clips using a computer vision algorithm that looks for scene changes. For computational reasons, we excluded all clips under 7 seconds. We were left with 4,537 clips, which totaled 52:24 hours.

Much of this footage was shot on the ground by protesters, bystanders, and photojournalists, while some was filmed from buildings or other aerial positions. The videographers tended to be dispersed over the space, and each one of the videographers is focusing on something slightly different. This is not surprising given the size and scope of the protests. These characteristics made audio synchronization the only reasonable approach in this case. Had the event been very focused and the location compact, visual synchronization may have been possible.

After creating a unique sound print for each clip using the algorithm described above that recognized a standardized vocabulary of features, we then compared the sequence of these features in each clip to all other clips and looked for reasonable matches. We had to include a certain amount of tolerance because of the challenges described above. We were able to synchronize 4:16:13 hours, some of which overlapped with the existing 4:37:19 manually synchronized hours. The combined total of synchronized video was increased by approximately 50 percent compared to the manually processed results.

Figure 2 shows an example of video clips synchronized by our system, displayed in a comprehensive, global view. The system also organizes all videos into a chronological timeline (shown in the upper panel of Figure 2). Researchers can access and play each video from the timeline. This example shows a person with green poncho (circled in red) who appeared at the same time in three of the videos. This person gives a visual evidence that the synchronization was successful.



**Figure 2. Synchronization of Conflict Videos (Global Time View).** A protester with a green poncho appears at the same time and place in three separate videos, providing visual evidence of correct synchronization. The other videos in this set show different locations at the same time.

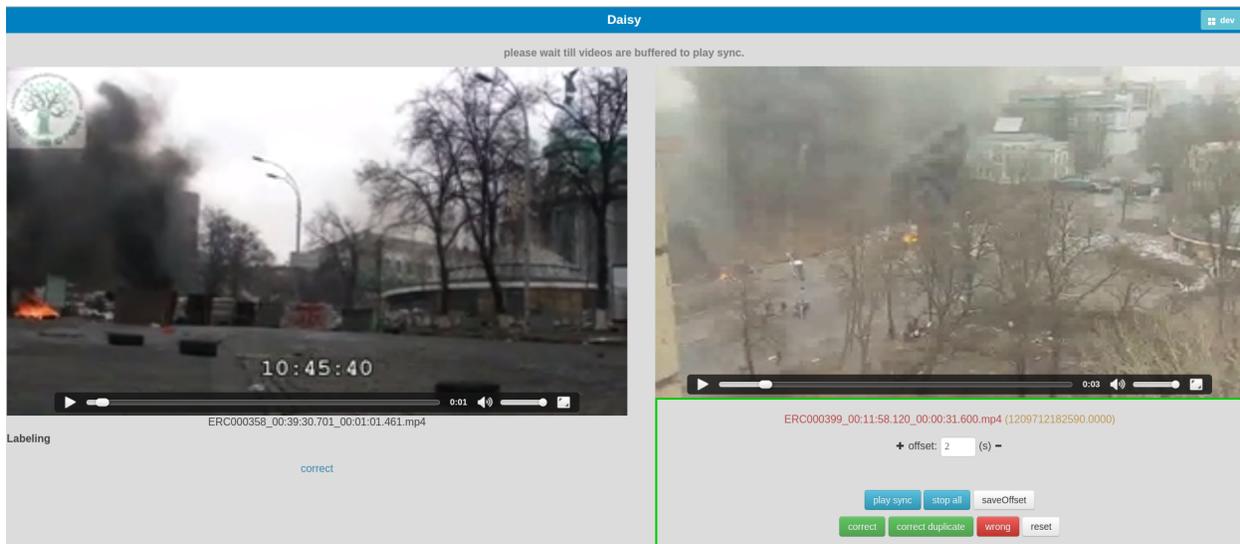
The synchronization of several videos is noisy and will likely include mistakes—this is precisely why human involvement in the process is crucial. To be able to review and

evaluate the results manually without getting overwhelmed by looking at several videos at the same time, we designed a browser that enables a human viewer to examine the synchronization results in an accurate pairwise view (see Figure 3).

There are several possible categories of matches or mismatches:

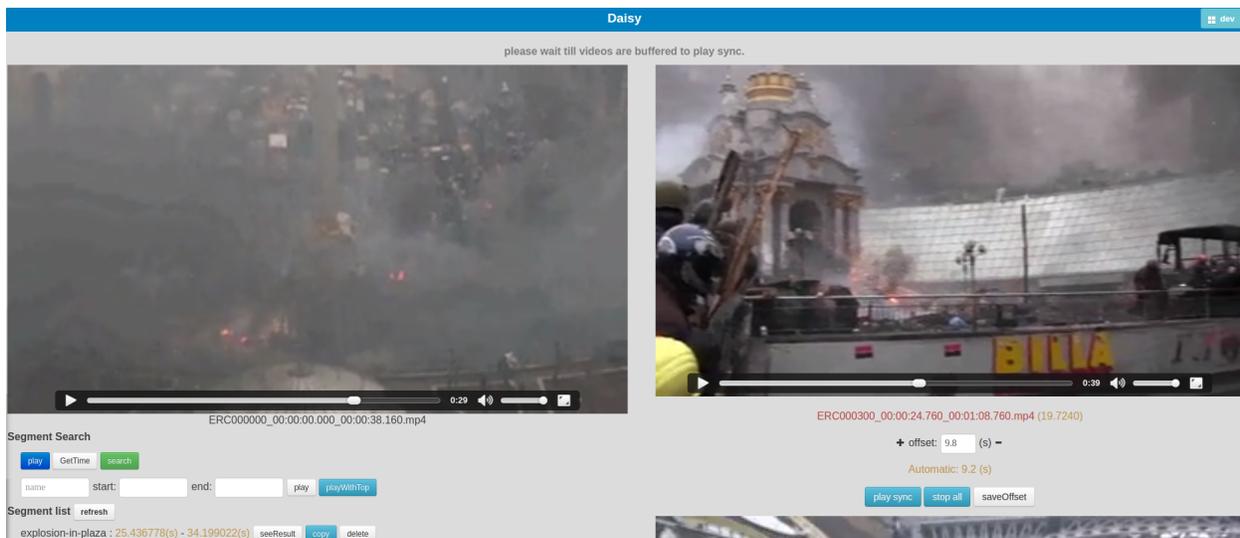
- the audio matches, but the video is different (i.e., the two videographers were standing near one another but were looking in opposite directions, or the audio captured a significant noise while the video was captured at different locations);
- the videos depict the same scene but are filmed from different points of view;
- the audio and the video are identical because the video segments are duplicates of each other;
- the match is a false positive, in that the audio seems similar to the system but is actually different.

If the matches are incorrect, or are exact duplicates of one another, the user can manually label them as such. Assuming that the two video segments are correctly synchronized, the researcher can manually adjust the offset determined by the system to correct for any differences in visual cues that are misaligned because of a camera's differing distances to the source of the sound. All of this information is saved and can be downloaded later for reference.



**Figure 3. Synchronization of Conflict Videos (Pairwise View)**

Our system can also perform a "segment-of-interest" search on videos. As shown in Figure 4, users can select a segment within the video containing the event they are interested in (for example, a series of explosions in a plaza), and search in other videos for a similar segment that shows similar looking buildings or persons, or that contains a similar sounding noise. A user may for example select a shooting scene with a significant series of gunshots, and may search for segments with a similar sounding series of gunshots. This method increases the chances for finding video scenes of an event displaying different angles of the scene or parallel events.

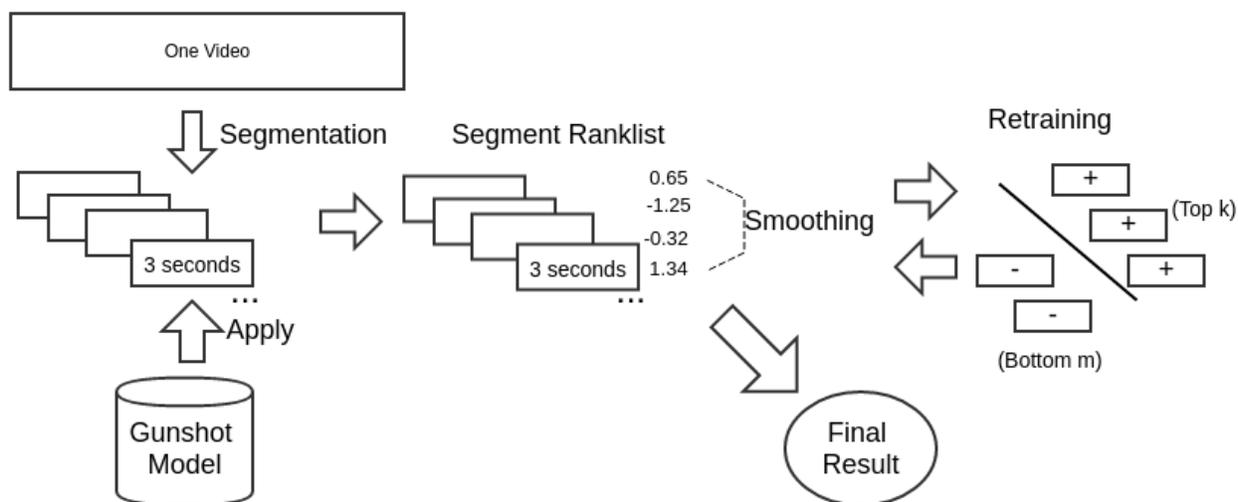


**Figure 4. Segment-of-Interest Search**

### Frame-level Detection of Specific Sounds

Our frame-level signature detection has been designed to use a similar algorithm as described above to detect audio events within a video - e.g., gunshots. Given a lengthy video - e.g., an hour-long video documenting a protest - finding the exact time when a gunshot happens manually will require a human reviewer to watch and listen to the entire video at least once at a reasonable playback speed. Our system aims to minimize that effort by automatically finding the segments within the video that have the highest probability of containing a specific audio event like gunshots.

This approach works by first segmenting a video into short snippets overlapping each other and by extracting the audio features from each of these snippets. Then a pre-trained classifier - in our example a “gunshot detector” (i.e., the spectral signature of sounds known to be gunshots) - filters through each of the snippets and produces a ranked list of those snippets that most likely contain gunshots. Finally, we assume the top snippets to be positive samples of gunshots and re-train a new gunshot detector using the new positive samples. The new gunshot detector filters through all the snippets again. The re-training process allows the system to learn what is extraneous noise in the video, and increases the number of positive gunshot samples. This extra step ultimately makes the system more accurate. Similar to the work on video synchronization, human investigators will then be able to quickly and efficiently verify whether a highly ranked snippet actually contains the queried gunshots. The frame-level detection pipeline is shown in Figure 5.

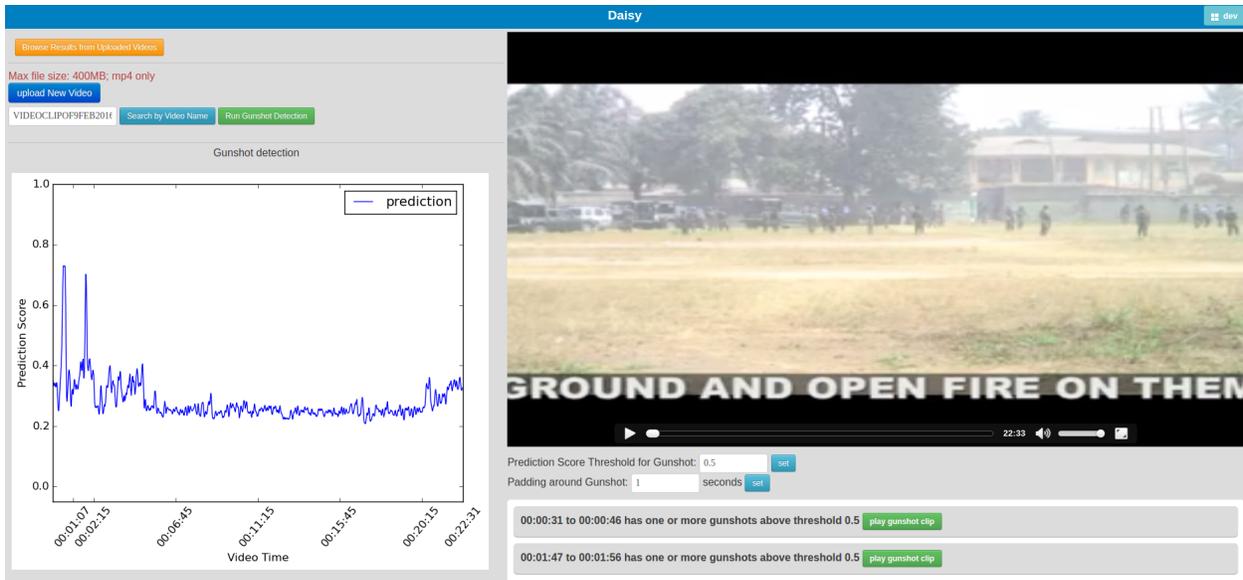


**Figure 5. Frame-level Sound Detection Using a Gunshot Classifier**

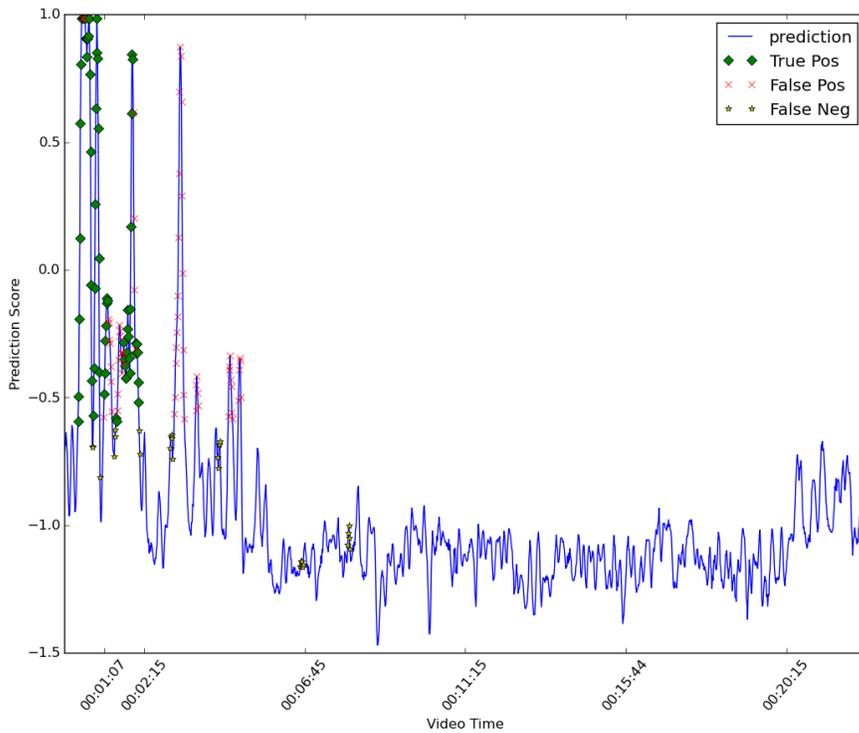
**Example: Gunshot Detection for Nigeria Protest Videos**

At the request of a partner organization that monitors rights violations in Nigeria, we were asked to see if we could count the number of gunshots in a 22-minute video depicting the violent breakup of a protest by armed police officers. After manually labeling all gunshots and running our system, it quickly became clear to us that our algorithm could not isolate individual gunshots, but could accurately detect gunshot activity at the frame level. In other words, we could tell the investigator where in the video most gunshots were likely to be occurring but could not give precise counts. Figure 6 provides a screenshot of the sound detector user interface.

As shown in Figure 7, the precision of gunshot detection on this video is 51% (i.e., more than half of the segments labeled as having gunshots actually did contain gunshots) and the recall is 69% (i.e., nearly 7 out of 10 actual gunshot events were detected by the system, while 3 out 10 were not). The system is not perfect, of course, but it will enable investigators pressed for time to more rapidly find gunshots in a video of interest. This algorithm can be trained to search for most well-defined sounds in addition to gunshots. It should be noted that we did not seek to push the limits of the technology in this test - only to show proof of concept. Subsequent refinement of the algorithm could likely increase its accuracy and reliability significantly if there is interest for us to do so in the human rights community.



**Figure 6. Gunshot Detection for Nigeria Protest Videos**



**Figure 7. Gunshot Detection Accuracy Graph.** *The green diamonds show segments that were correctly labeled as containing gunshots. The red Xs show areas that the system incorrectly labeled as containing gunshots, and the grey stars show segments containing gunshots that were missed by the system. The blue line tracks the probability of correctness that the system attaches to each segment. The closer to 1.0, the more likely the segment is to have gunshots according to the trained classifier.*

## Conclusion

The machine learning-based audio analysis system described in this technical report can help human rights practitioners synchronize large volumes of audio-rich video, and search for specific sounds within their video collections. The goal of our system is to make synchronization more manageable for human analysts by narrowing down the amount of video that needs to be closely examined. Our system does not eliminate human involvement in the process because machine learning systems provide probabilistic, not certain, results. Synchronization is only the first step in the analysis of large video collections. In subsequent publications, we will describe the tools and methods we are developing to organize and visualize this content.

## References

1. Sameer Padania, Sam Gregory, Yvette Alberdingk-Thijm, and Bryan Nunez, "Cameras Everywhere: Current Challenges and Opportunities at the Intersection Of Human Rights, Video and Technology," [http://www.ohchr.org/Documents/Issues/Opinion/Communications/Witness\\_1.pdf](http://www.ohchr.org/Documents/Issues/Opinion/Communications/Witness_1.pdf) (Witness, 2011).
2. Sam Dubberley, Elizabeth Griffin, and Haluk Mert Bal, "Making Secondary Trauma a Primary Issue: A Study of Eyewitness Media and Vicarious Trauma on the Digital Frontline," <http://eyewitnessmediahub.com/research/vicarious-trauma> (Eyewitness Media Hub, 2015).