

Gun Source and Muzzle Head Detection

Zhong Zhou¹, Isak Czeresnia Etinger¹, Florian Metzger¹, Alexander Hauptmann¹, Alexander Waibel^{1, 2}

¹ Carnegie Mellon University, Pittsburgh, PA, the USA

² Karlsruhe Institute of Technology, Karlsruhe, Germany

Abstract

There is a surging need across the world for protection against gun violence. There are three main areas that we have identified as challenging in research that tries to curb gun violence: temporal location of gunshots, gun type prediction and gun source (shooter) detection. Our task is gun source detection and muzzle head detection, where the muzzle head is the round opening of the firing end of the gun. We would like to locate the muzzle head of the gun in the video visually, and identify who has fired the shot. In our formulation, we turn the problem of muzzle head detection into two sub-problems of human object detection and gun smoke detection. Our assumption is that the muzzle head typically lies between the gun smoke caused by the shot and the shooter. We have interesting results both in bounding the shooter as well as detecting the gun smoke. In our experiments, we are successful in detecting the muzzle head by detecting the gun smoke and the shooter.

Introduction

There is a surging need across the world for protection against gun violence. There are 17,502 gun violence incidents that resulted in 4606 deaths in the United States alone up to date; among which 105 are mass shooting as shown in the Gun Violence Archive¹. Mass shooting is defined by the Mass Shooting Tracker² as any incident that involves shooting 4 people or more in one incident independent of any circumstance.

From the 2017 Las Vegas shooting where rows of semi-automatic machine guns were pointed at innocent people, to the 2017 Texas Sutherland church shooting where many were killed during Sunday worship; from the 2016 nightclub shooting where many members of the gay community were killed, to the 2018 Pittsburgh Tree of Life synagogue shooting where a lot of animosity were shown towards innocent members of the Jewish community. The loss of life, the contribution to group segregation, tension, division and conflict, and the damage done to the survivors and to each community targeted is immeasurable. Therefore, there is an insurmountable and tremendous need for curbing gun violence for humanitarian efforts³.

Many researchers are interested in contributing to the efforts to mitigate gun violence. There are three main areas that we have identified as challenging: temporal location of gunshots, gun type prediction and source detection. Temporal location of gunshots in essence is finding the probability distribution of the likelihood of a



Figure 1. Tree of Life synagogue [1].

gunshot to be found in a temporal segment of the audio. It is very important to extract accurate signal at this stage for subsequent research. Gun type detection involves detection and discerning which type of gun is involved. Gun source detection involves finding out who fires the shot, especially in the scenario where there are multiple people in the video frame. And there may be multiple people carrying guns in the video frame but not everyone is shooting. There are many researchers who focused on the first two problems and achieved good results through Localized Self-Paced Reranking by [3, 4, 5, 6, 7, 8, 9].

Our task is gun source detection and muzzle head detection. Muzzle head is the round opening of the firing end of the gun. We would like to visually locate the muzzle head of the gun in the video, and identify who has fired the shot. This is a very difficult task. There are very few real-life events that contain the footage of the shooter, compared to a multitude of videos of the victims. Even in situations where both gun and shooter are visible, it is difficult to visually detect whether there is indeed a gunshot, it is therefore even more difficult to discern who fires it. It usually takes a long time to manually detect who fires the shot, and in many cases, even manual detection would fail. There is a huge gain in automating the process, or at least in greatly facilitating human efforts on detecting who fires the shot and locate the muzzle head where a shot is fired.

Related Work

Many researchers have worked on gunshot prediction as well as gun type detection using audio data. The first wave of work with this strategy was performed between 2005 and 2013.

Researchers detected abnormal audio events in continuous audio recordings of public places and focused on robustness and prioritized recall over precision [10]. Some worked on the modelling of gunshot trajectories by simple geometry and kinematics, using the time taken for sound to travel from a gun to a recorder as

¹<https://www.gunviolencearchive.org>

²<https://www.shootingtracker.com>

³Images of various shooting events in this paper are taken from <https://www.wcjb.com>, <https://www.chabad.org>, <https://www.nbcnews.com>, and <https://www.bbc.com/>

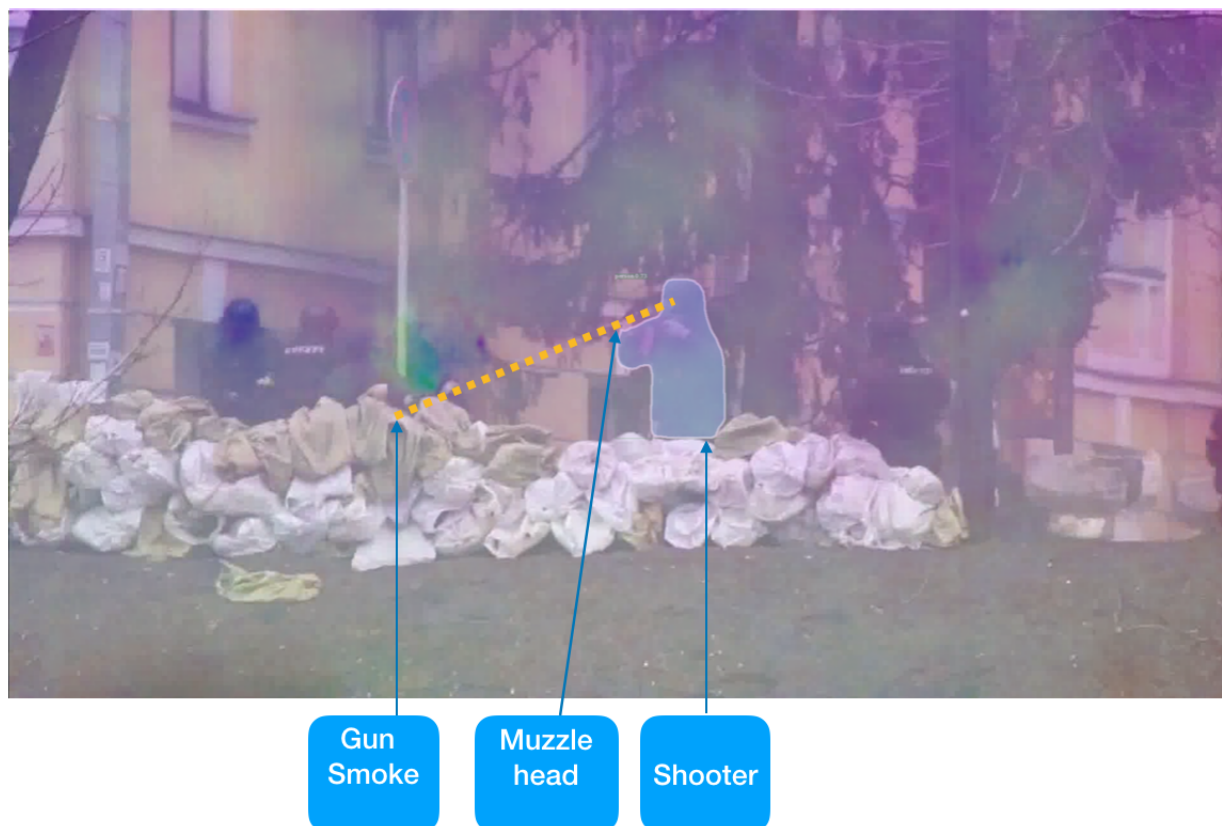


Figure 2. Overlay of optical flow visualization over the original video with human detection [2].

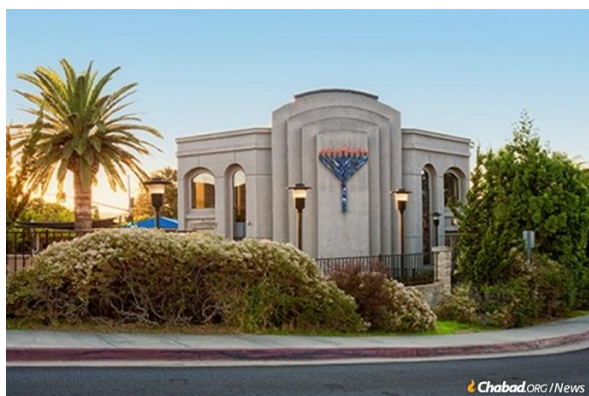


Figure 3. Chabad of Poway Synagogue (near San Diego) [17].

an indication of distance from the recorder [11]. Some expanded on the work of the two previous papers by building a system using two Gaussian Mixture Model classifiers for detecting gunshots and screams, and also using kinematics to model gunshot trajectories [12, 13]. Some used dynamic programming and Bayesian networks to detect gunshots from audio streams from movies [14]. Many extended evaluation of previous techniques to extremely noisy environments, by recording gunshots at open fields and adding white noise [15]. The efficiency of several methods from 2005 to 2010 is summarized by [16].

In 2012, Kumar’s team published a work that automatically learned atomic units of sound, called Acoustic Unit Descriptors [23]. In 2013, Ahmed formulated a two-stage approach that uses an event detection framework followed by a gunshot recognition stage, which used a template matching measure and the eighth order linear predictive coding coefficients to train an SVM classifier [24].

A second wave of more advanced methods came later [25, 3, 26, 4]. Liang created a tool that synchronizes multiple videos of a single event (especially event involving gunshots) by creating a unique sound signature at the frame level for each video in a collection [25]. Liang worked on the temporal localization of gunshots, and employed Localized Self-Paced Reranking (LSPaR) to refine the localization results [3, 27]. LSPaR utilizes curriculum learning (i.e.: samples are fed to the model from easier to noisier ones) so that it can overcome the noisiness of the initial retrieval results from user-generated videos. Additionally, SPaR has a concise mathematical objective to optimize and useful properties that can be theoretically verified, instead of relying mainly on heuristic weighting as other reranking methods. Lim introduces a rare sound event detection system using a 1D convolutional neural network and long short term memory units (LSTM). MFCCs are used as input; the 1D ConvNet is applied in each time-frequency frame to convert the spectral feature; then the LSTM incorporates the temporal dependency of the extracted features [26].

Audio-based gunshot detection and temporal localization have

Video id	Smoke color+ intensity[1-5]	Background color	Video resolution	Camera far?	Gun stable?	Shooter moves?	Camera moves?	Gun position w.r.t. camera	Shot/shooter obstruction	Shooter pose
1	grey, 5	grey	good	no	no	no	no	pointed up	people	standing
2	grey, 2	grey	medium	no	yes	no	no	sideways	nothing	standing
3	grey, 1	grey	bad	no	yes	no	no	sideways	nothing	kneeling
4	orange, 5	white	bad	no	no	yes	no	sideways	nothing	kneeling
5	grey, 1	white	medium	no	yes	no	no	sideways	nothing	lying
6	orange, 5	white	bad	no	no	yes	no	behind	nothing	standing
7	grey, 2	grey	medium	no	yes	no	no	sideways	nothing	standing
8	grey, 1	grey	medium	no	yes	no	no	sideways	nothing	lying
9	grey, 1	grey	medium	no	no	yes	no	sideways	tree	walking
10	grey, 1	grey	medium	no	no	yes	no	sideways	tree	walking
11	orange, 1	grey	medium	no	yes	no	no	sideways	nothing	kneeling
12	orange, 4	grey	medium	no	yes	no	no	sideways	nothing	kneeling
13	grey, 1	grey	bad	yes	yes	no	no	sideways	nothing	standing
14	grey, 2	grey	good	no	yes	no	yes	sideways	tree	standing
15	grey, 3	grey	medium	no	no	yes	yes	sideways	nothing	walking

Statistic of videos with visible gunshots.

many applications. Some are patents of real products [28, 29]. Aronson developed a tool to monitor conflicts [30]. This is because audio is often the only source available. For gun type detection, clean video gunshots with very explicit source like the well-known Eddie Adams' Vietnam war picture (Figure 5) are very rare. In general terrorist attacks, it is very unlikely to have people recording data because attacks are unsuspected (so people would have no reason to be filming the action beforehand), and because people seek cover as soon as the threat is realized. For example, Chen showed that the vast majority of recorded data of the Boston Marathon terrorist attack was not directly aimed at the action [31].

However, with the rapid increase in the use of smart phones with high-resolution cameras in the past several years, there has been a growing opportunity to leverage video data for gunshot detection and localization. As most of the analyzed video data of terrorist attacks or human-rights violations does not usually involve gunshots, it is hard to find useful real-life data for research. Even in the few videos where gunshots are observed, the guns are usually not visible or extremely small, this renders the task to be very challenging. Take the well-known Eddie Adams' Vietnam war picture for example again, it is very rare that we have the shooter, the gun, and the victim in the same picture. For the 2017 Las Vegas shooting, thousands of videos are on the screaming crowd of victims, but there is no footage about the shooter, because the shooter is shooting from a highly elevated hotel room using rows of semi-automatic machines guns that are invisible from any visual



Figure 5. Execution of Nguyễn Văn Lém, by Eddie Adams.

sources. Indeed, videos with complete information of the guns are very rare.

There is much research done on image-based gun detection [32, 33, 34, 35]. Zhang was one of the pioneers in image-based gun detection. He worked on region-based image fusion for concealed weapon detection [34]. Sun then worked on the detection of gun barrels by using segmentation of forward-looking infrared (FLIR) images with fuzzy thresholding and edge detection [32]. Xue compared the performances of a large set of image fusion algorithms for concealed weapon detection using visual and infrared images [33]. Finally, Tiwari proposed a framework that exploits the color based segmentation to eliminate unrelated object from an image using k-means clustering, then Harris interest point detector and Fast Retina Keypoint is used to locate the guns in the segmented images [35].

But still, this is done using only images – not complete video

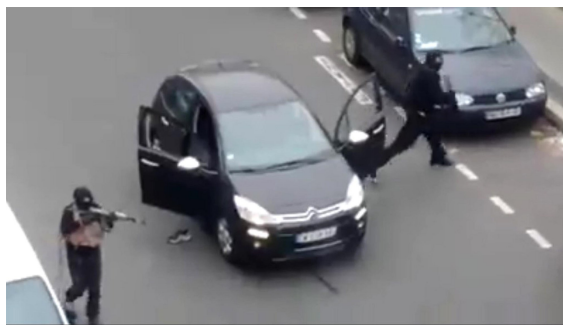


Figure 4. Charlie Hebdo terrorist shooting [18].

Gun Cloud Detection Success rate	Shooter Detection Rate	Muzzle Head Detection Rate
69.6%	30.4%	21.7%

Performance of gun smoke, shooter, muzzle head Detection.



Figure 6. Results from human object detection mechanism through the Detectron network [19].

data. To the best of our knowledge, there is no existing literature that uses video data as well as audio data together. The temporal localization of gunshots and the gun type detection are very well done, mainly based purely on sound [3, 26, 4, 9].

However, gunshot source detection is still very untackled, and this work seems to leverage the combination of sound and video to address that challenge. We differentiate from previous research not only by using both visual and sound data jointly, but also by using video features instead of only static image features.

Data

We have four datasets: Urban Conflict, Target Range, Urban Sound, and Real-World Events.

Urban Conflict

The Urban Conflict dataset consists of 422 videos of conflicts in Ukraine gathered by news sources. The videos are split into 4537 segments, and they sum to 52.5 hours in total duration. The average length of each video is therefore 7.5 minutes.

Target Range

The Target Range dataset consists of 547 soundtracks of gunshots downloaded from YouTube videos of target range practices, further divided into 3761 segments [3].

Urban Sound

The Urban Sound dataset consists of 1302 audio files of field recordings [36]. This dataset provides sound on a wide range of noises common to the urban environment. We used only the gunshot data from this dataset. Systematic urban sound classification is a new field of research with many applications. There is scarcity of a common taxonomy.

Real-World Events

The Real-World Event dataset was gathered from recordings of several different events: Las Vegas Shooting, Santa Fe School shooting, Orlando Night Club shooting, Douglas High School shooting, Jacksonville Tournament shooting, Dallas shooting, Florida school shooting, Thousand Oaks shooting, Sutherland Springs church shooting, Kansas taser shooting [9].

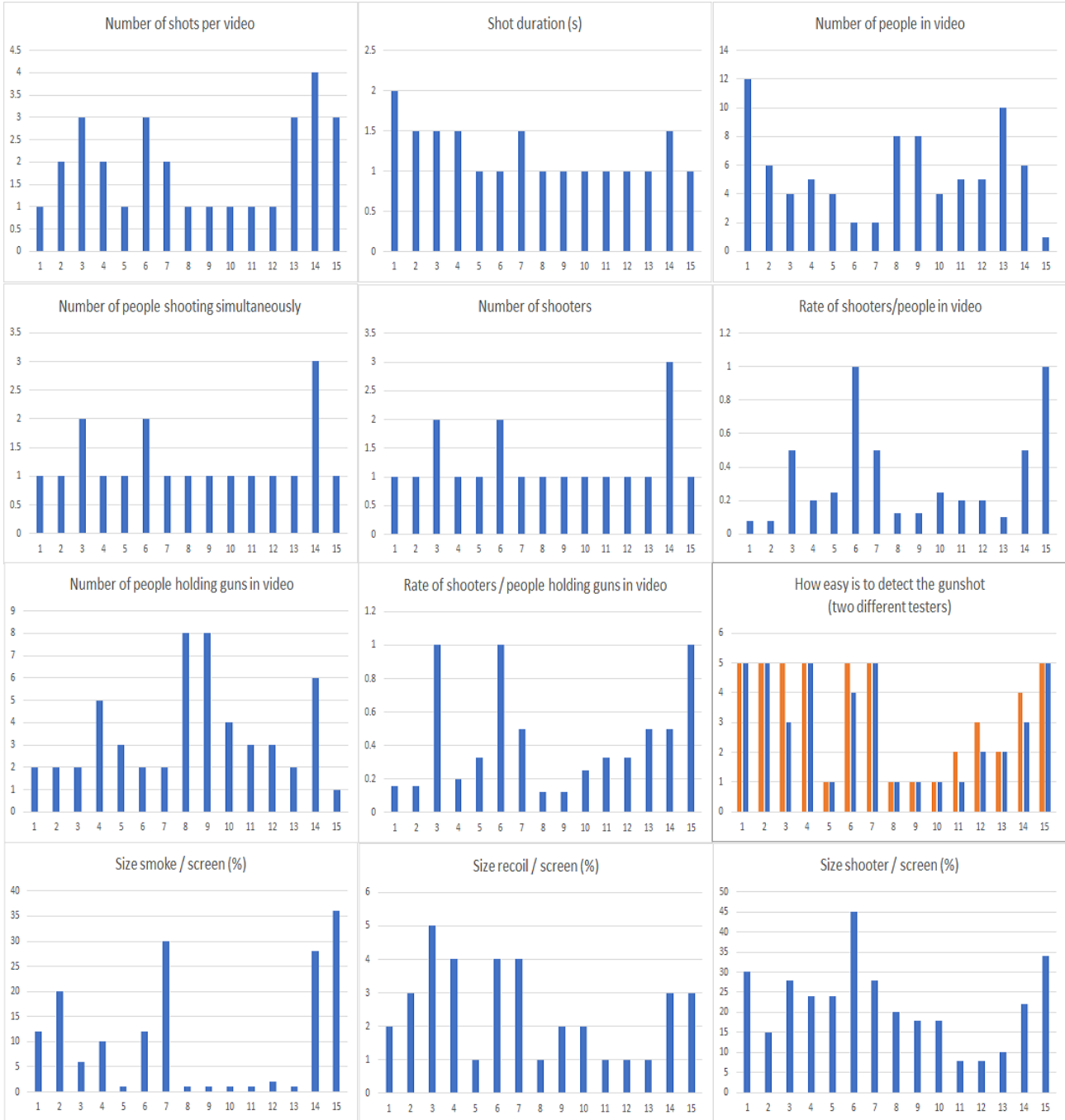


Figure 7. Statistics of videos with visible guns. The horizontal axis is the video id.

Experiments, Methods, and Data Flow

Gunshot Probability Distribution through Localized Self-Paced Reranking

We applied the model from Liang on the Urban Conflict dataset [3]. The process consists of three steps:

1. First the audio stream is extracted from the videos and chunked into small segments of 3-second windows with a 1 second stride. Bag-of-Words of MFCC features is employed to represent each segment [37, 38].
2. Then, two-class SVM classifiers are trained for each audio event and applied to the video segments from test videos. However, the initial detection results have low accuracy due to noise.
3. After the detector model produces an initial ranked list of video segments, we utilize LSPaR to learn a reranking model with curriculum learning (first "cleaner" videos with high confidence scores are fed, then "noisier" videos with smaller confidence scores).

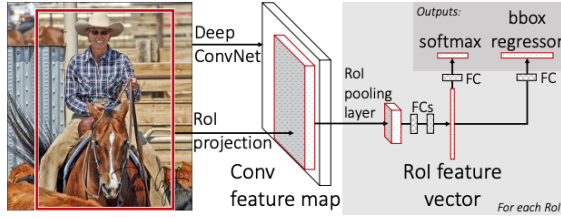


Figure 8. fast R-CNN architecture [20].

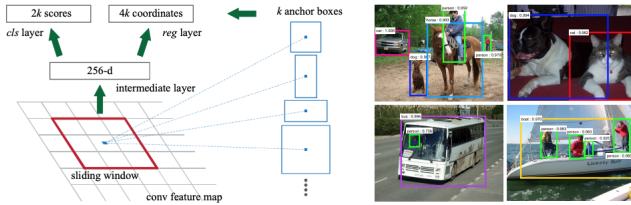


Figure 9. Faster R-CNN's use of RPN [21].

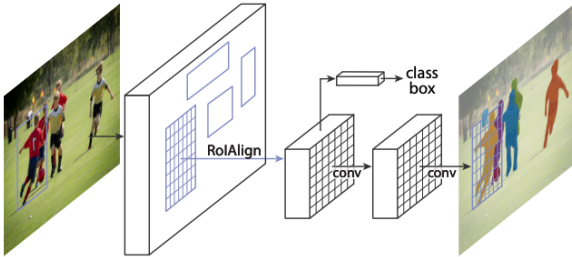


Figure 10. Mask R-CNN architecture [2].

This produced a list of 3-second video segments ranked by a confidence score of how likely there is a gunshot in the video, based purely on audio information.

Manual Selection of Videos with Visible Shooting

Once we obtained the probability distribution of possible gunshots in each video segment, we choose the video segments with the probability of gunshot (as calculated in the previous Subsection) exceeding a threshold of 70%. After this step, we manually watched all videos to find the videos that contain visible gunshots.

We found 15 non-overlapping small segments of videos that contained a visible gunshot. Many of those videos were extremely hard to find manually. The reasons are the low resolution of the videos, the small size of the gun recoil and the smoke relative to the screen size; in addition, there are typically many hundreds of hours of videos to be analyzed before a gunshot is observed. Several statistics of those 15 videos can be found on Table and in Image 7.

Human Object Detection Through Detectron

After the audio temporal localization of gunshots and the manual search for videos with visible shooting scenes, we obtained a set of 15 non-overlapping small segments of videos that contained a visible gunshot, and many had multiple gunshots happening concurrently. After examining the videos closely, we realize that the camera is moving in many of the videos. As the camera moves,

all objects in the frame moves. Therefore, it is hard to track a fixed point like the muzzle head automatically.

Our idea is to track the shooter instead of tracking the muzzle head. There are several reasons for this. Firstly, the shooter is much bigger than the muzzle head, instead of doing small (tiny) object detection, we can do usual object detection, which is researched on more extensively by the scientific community. Secondly, there are existing open-source models trained on people detection, but not on muzzle head because there are very few instances of muzzle head in most of the image datasets. Thirdly, even though the camera may move, the shooter and the muzzle head are relatively static with respect to each other as the muzzle head is in the direction of the shooter's eyes and is in the direction where the shooter is facing.

In order to detect people in images, we use Facebook's implementation of Mask R-CNN, Detectron [19]. Fast R-CNN has achieved huge success in the visual domain as shown in Figure 8 [20]. It has advantages over the state-of-the-art object detection benchmark R-CNN and SPPnet in that it has much better mAP scores (better detection accuracy) while training in the single-stage fashion utilizing multi-task loss function. Every network layer is updated concurrently at the same time during training while no extra disk space is needed during each stage of feature caching.

Faster R-CNN builds on Fast R-CNN by presenting a Region Proposal Network (RPN) as shown in Figure 9 [21]. RPN enables sharing of CNN features with the network that does object detection facilitating low-cost region proposal scheme.

Mask RCNN carries Faster R-CNN further by adding an object mask prediction on top of the bounding box prediction as shown in Figure 10. We use Facebook's implementation of Mask RCNN to detect people in our videos [19].

Human Evaluation of People Detection

For most videos, our results of human detection are good; however, when videos have low-resolution or when there are obstructions between the camera man and the shooter(s), it is hard for the human detection mechanism to detect all the shooters involved as shown in Figure 6. We also find that if a shooter lies flat on the ground, it is hard for detectron to detect this shooter. This is a very interesting problem because people, unlike guns or muzzle head, are deformable; a person is capable of having different shooting poses as shown in Table [39]. Most of the shooters shoot behind some barriers, either standing or squatting down, and very few



Figure 11. Las Vegas shooting [22].



Figure 12. Comparison between original video frame and its respective Optical Flow graph. Left: Video keyframe with Visible Shooting. Right: Optical Flow Graph showing clearly the shooting smoke.

training data involves shooters lying flat on the ground. One of the hardest pose of a shooter is the pose of shooting while lying on the ground. In human detection training data, the majority of the data samples involves people standing, sitting, it is extremely sparse to find training data where a person is lying flat on the ground. Therefore, identifying shooters lying on the ground is challenging.

Optical Flow, and Flownet 2.0

In order to detect visible shooting cloud, we use Flownet 2.0, which is the evolved estimation of Optical Flow algorithm using deep Neural Networks [40]. Flownet formulated Optical Flow estimation in deep Convolutional Networks as a supervised learning problem [41]. Flownet 2.0 enhanced Flownet to cover minor movements and noisy real-world data by adding a small sub-network that covers minor movements and utilizing a stacked network scheme [40]. We employ flow2image⁴ to visualize our

⁴Available at <https://github.com/georgegach/flow2image>, originally based on Daniel Scharstein (C++) and Deqing Sun (MATLAB)'s

outputs from Flownet 2.0. In Figure 12, the right hand side contains the flow visualization.

Human Evaluation of flow visualizations

In Figure 12, we show a few examples of our results. The left hand side contains original shooting footage, and the right hand side contains the flow visualization encapsulating the change from this shooting keyframe to the next keyframe. For simplicity, we show only the keyframe before the flow is observed and not the keyframe after.

In the first set of examples, we see a shooter shooting to the left. It is hard for human to detect gunshot visually without listening to the gunshot because the white smoke fades into the background which is also grey and white. However, our optical flow graph clearly shows a blue cloud with very static background. This is a clear signal of a gunshot.

In the second set of examples, we have a very low-resolution work <http://vision.middlebury.edu/flow/>



Figure 13. People Gathered to Pray for Victims' Families [42].

video that any human would find very difficult to tell what is inside the darkness; a few people with practiced eyes or with military training may detect gunshots visually. However, this is very hard, neither people nor gun are visually apparent. When we turn to the flow graph, we see very clear a blot of purple cloud which is a very clear signal of gunshot.

In the last example, the same level of visual obscurity is observed, the entire video is grey instead of dark, and what makes this video harder is the angle of the camera is from above and far away from the shootings. There are also many people in the video. However, in the flow visualization, we see a very clear and concentrated red dot with a tint of green, signally gunshot clearly.

Conclusion

We focus on gun source detection and muzzle head detection in the hope of mitigating gun violence and helping the police to detect the severity of a gun shooting incident as fast as possible. The huge need across the world for protection against gun violence is our main motivation and we as researchers hope to contribute our share to world peace and social stability.

In our research, we turn our problem formulation to make use of human object detection and gun smoke detection and have interesting results both in bounding the shooter(s) as well as detecting the gun smoke. The muzzle head is found between the gun smoke and the shooter.

Indeed, when we overlay the optical flow visualization with the human detection output, we see clearly where the muzzle head is. The muzzle head is between the gun smoke on the left and the shooter on the right. We label the shooter, muzzle head and gun smoke as shown in Figure 2. Our results for gun smoke detection are much higher than for shooter detection; in Table 2, our evaluation is done using the 15 videos having visible gunshots, gun cloud detection's success rate is 69.6% while shooter human detection's success rate is 30.4%. Since muzzle head detection (21.7%) relies on both gun cloud detection and shooter detection, the success rate of muzzle head detection has a great potential to improve, and benefits the most if we can improve the shooter human detection rate as shown in Table 2.

In the future, researchers may attempt to locate the exact location of the muzzle head in the direction of the gun smoke. The muzzle head should be on the straight line connecting the center of the gun cloud and the shooter's eye. However, though it is relatively easier to detect people, it is hard to locate a person's eyes. Also, though the gun cloud is visible and in most cases

round, sometimes it is oval or irregular. Therefore, it may be hard to identify the center of the cloud. But all of these are interesting questions for future research.

Acknowledgments

We would like to thank Carla De Oliveria Viegas, Siddhartha Sharma, Junwei Liang, Ruonan Liu and Ankit Shah for their valuable feedbacks and their generous sharing of ideas, data, models, and feature files.

References

- [1] WCJB, “Tree of life synagogue,” 2019. [Online; accessed Mar 1, 2019].
- [2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- [3] J. Liang, L. Jiang, and A. Hauptmann, “Temporal localization of audio events for conflict monitoring in social media,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1597–1601, IEEE, 2017.
- [4] L. Ruonan, “Unpublished.” 2019.
- [5] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Transactions on information theory*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [6] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.
- [7] R. R. Coifman, Y. Meyer, and V. Wickerhauser, “Wavelet analysis and signal processing,” in *In Wavelets and their applications*, Citeseer, 1992.
- [8] H. Zhang, N. D. Daw, and L. T. Maloney, “Human representation of visuo-motor uncertainty as mixtures of orthogonal basis distributions,” *Nature neuroscience*, vol. 18, no. 8, p. 1152, 2015.
- [9] S. Ankit, “Unpublished.” 2019.
- [10] C. Clavel, T. Ehrette, and G. Richard, “Events detection for an audio-based surveillance system,” in *2005 IEEE International Conference on Multimedia and Expo*, pp. 1306–1309, IEEE, 2005.
- [11] R. C. Maher, “Modeling and signal processing of acoustic gunshot recordings,” in *2006 IEEE 12th Digital Signal Processing Workshop & 4th IEEE Signal Processing Education Workshop*, pp. 257–261, IEEE, 2006.
- [12] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 21–26, IEEE, 2007.
- [13] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection in noisy environments,” in *2007 15th European Signal Processing Conference*, pp. 1216–1220, IEEE, 2007.
- [14] A. Pirkakis, T. Giannakopoulos, and S. Theodoridis, “Gunshot detection in audio streams from movies by means of dynamic programming and bayesian networks,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 21–24, IEEE, 2008.
- [15] I. L. Freire and J. A. Apolinario Jr, “Gunshot detection in noisy environments,” in *Proceeding of the 7th International Telecommunications Symposium, Manaus, Brazil*, pp. 1–4, 2010.
- [16] A. Chacon-Rodriguez, P. Julian, L. Castro, P. Alvarado, and N. Hernández, “Evaluation of gunshot detection algorithms,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 2, pp. 363–373, 2011.
- [17] “Chabad of poway synagogue.” https://www.chabad.org/news/article_cdo/aid/4365316/jewish/One-Dead-Three-Injured-in-Shooting-at-Chabad-Synagogue-Near.html.
- [18] “Charlie hebdo terrorist shooting.” <https://www.nbcnews.com/storyline/paris-magazine-attack/charlie-hebdo-shooting-12-killed-muhammad-cartoons-magazine->
- [19] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, “Detectron.” <https://github.com/facebookresearch/detectron>, 2018.
- [20] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.
- [21] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, pp. 91–99, 2015.
- [22] “Las vegas shooting.” <https://www.nbcnews.com/storyline/las-vegas-shooting/las-vegas-police-investigating-shooting-mandalay-bay-n80646>
- [23] A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, and B. Raj, “Audio event detection from acoustic unit occurrence patterns,” in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 489–492, IEEE, 2012.
- [24] T. Ahmed, M. Uppal, and A. Muhammad, “Improving efficiency and reliability of gunshot detection systems,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 513–517, IEEE, 2013.
- [25] J. Liang, S. Burger, A. Hauptmann, and J. D. Aronson, “Video synchronization and sound search for human rights documentation and conflict monitoring,” 2016.
- [26] H. Lim, J. Park, and Y. Han, “Rare sound event detection using 1d convolutional recurrent neural networks,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, pp. 80–84, 2017.
- [27] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, “Easy samples first: Self-paced reranking for zero-example multimedia search,” in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 547–556, ACM, 2014.
- [28] K. C. Baxter and K. Fisher, “Gunshot detection sensor with display,” Sept. 4 2007. US Patent 7,266,045.
- [29] V. Cowdry, “Gun shot detector,” Dec. 11 2014. US Patent App. 14/300,771.
- [30] J. D. Aronson, S. Xu, and A. Hauptmann, “Video analytics for conflict monitoring and human rights documentation,” *Center for Human Fights Science Technical Report*, 2015.
- [31] J. Chen, J. Liang, H. Lu, S.-I. Yu, and A. G. Hauptmann, “Videos from the 2013 boston marathon: An event reconstruction dataset for synchronization and localization,” 2016.
- [32] S. Sun and H. W. Park, “Segmentation of forward-looking infrared image using fuzzy thresholding and edge detection,” *Optical Engineering*, vol. 40, no. 11, pp. 2638–2646, 2001.
- [33] Z. Xue, R. S. Blum, and Y. Li, “Fusion of visual and ir images for concealed weapon detection,” in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002.(IEEE Cat. No. 02EX5997)*, vol. 2, pp. 1198–1205, IEEE, 2002.
- [34] Z. Zhang and R. S. Blum, “Region-based image fusion scheme for concealed weapon detection,” in *Proceedings*

of the 31st annual conference on information sciences and systems, pp. 168–173, 1997.

- [35] R. K. Tiwari and G. K. Verma, “A computer vision based framework for visual gun detection using harris interest point detector,” *Procedia Computer Science*, vol. 54, pp. 703–712, 2015.
- [36] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, ACM, 2014.
- [37] Q. Jin, P. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, “Event-based video retrieval using audio,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [38] J. Liang, Q. Jin, X. He, G. Yang, J. Xu, and X. Li, “Detecting semantic concepts in consumer videos using audio,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2279–2283, IEEE, 2015.
- [39] D. J. Moore, I. A. Essa, and M. H. Hayes, “Exploiting human actions and object context for recognition tasks,” in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 1, pp. 80–86, IEEE, 1999.
- [40] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017.
- [41] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, “Flownet: Learning optical flow with convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2758–2766, 2015.
- [42] “People gathered to pray for victims’ families.” <https://www.bbc.com/news/world-us-canada-46002549>.