Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Demonstrations Track

CVAT-BWV: A Web-Based Video Annotation Platform for Police Body-Worn Video

Parsa Hejabi[1] , **Akshay Kiran Padte**[1] , **Preni Golazizian**[1] , **Rajat Hebbar**[1] , **Jackson Trager**[1] , **Georgios Chochlakis**[1] , **Aditya Kommineni**[1] , **Ellie Graeden**[2] , **Shrikanth Narayanan**[1] , **Benjamin A.T. Graham**[1] and **Morteza Dehghani**[1]
[1]University of Southern California
[2]Georgetown University
{hejabi, padte, golazizi, rajatheb, jptrager, chochlak, akommine, shri, benjamag, mdehghan}@usc.edu, ellie.graeden@georgetown.edu

Abstract

We introduce an open-source platform for annotating body-worn video (BWV) footage aimed at enhancing transparency and accountability in policing. Despite the widespread adoption of BWVs in police departments, analyzing the vast amount of footage generated has presented significant challenges. This is primarily due to resource constraints, the sensitive nature of the data, which limits its widespread access, and consequently, lack of annotations for training machine learning models. Our platform, called CVAT-BWV, offers a secure, locally hosted annotation environment that integrates several AI tools to assist in annotating multimodal data. With features such as automatic speech recognition, speaker diarization, object detection, and face recognition, CVAT-BWV aims to reduce the manual annotation workload, improve annotation quality, and allow for capturing perspectives from a diverse population of annotators. This tool aims to streamline the collection of annotations and the building of models, enhancing the use of BWV data for oversight and learning purposes to uncover insights into police-civilian interactions.

1 Introduction

High-quality policing is an essential government service, yet it often unevenly impacts different communities [California Department of Justice, 2024]. In the US, police conduct over 20 million traffic stops annually, where the officers' communication, including their tone of voice and level of respect, shapes public perceptions of, and trust in, the police [Camp *et al.*, 2021]. The vast majority of these traffic stops are captured on body-worn video (BWV). Unfortunately, even large police departments currently have limited ability to analyze this data. Instead of transforming policing, BWV data often serves merely as an ad hoc crisis management tool, with footage rarely reviewed unless a complaint is filed or a tragic event occurs.

Machine learning (ML) tools have the potential to enable police departments and those tasked with their oversight to analyze BWV footage at scale, thus facilitating transparency and accountability. Relatively small-scale studies of BWV footage have already revealed significant racial disparities in officer-civilian interactions [Voigt *et al.*, 2017; Rho *et al.*, 2023; Prabhakaran *et al.*, 2018], highlighting the potential of BWV footage if analyzed effectively. However, developing ML models in the policing domain requires extensive human annotation of BWV recordings by a diverse population of annotators, not least because the perception of effective communication can differ widely between individuals and communities [Graham *et al.*, 2024]. These human annotations can help distinguish between speakers in overlapping dialog, track the progression of conversations, identify the (de)escalation of conflict, analyze communication styles from gestures to tone of voice, and also capture potential differences in viewpoints of groundtruths.

Several annotation tools have been developed to enhance the annotation process for multimodal ML tasks (e.g., Labelbox, Prodigy); however, none offer the comprehensive functionalities required for annotating BWV data. We have based our framework on Computer Vision Annotation Tool (CVAT) [Sekachev *et al.*, 2020], as it is an open-source annotation software that natively supports object detection and image classification. Moreover, it offers the flexibility of local hosting and the customizability inherent in open-source programs. Other existing tools, such as Labelbox [Labelbox, 2024] and Prodigy [Montani and Honnibal, 2024], were not suitable for our task; they either operate as Software as a Service platforms, which do not meet the security criteria for sensitive data like BWV, are incapable of handling all modalities (text, audio, and video), or are not sufficiently open-source and customizable to incorporate ML models for assisted annotation and enable modification of the user interface.

We introduce CVAT-BWV[1], an open-source multimodal interactive annotation platform for BWVs with the following features: 1) Secure and local processing of data, due to the sensitive nature of BWV footage; 2) Integration of AI for assisted annotation in various tasks, including object detection, automatic speech recognition (ASR), speaker diarization, and face recognition; and 3) Support for multimodal, and multi-perspective, annotations.

[1]https://www.youtube.com/watch?v=LyfLsG8OQug

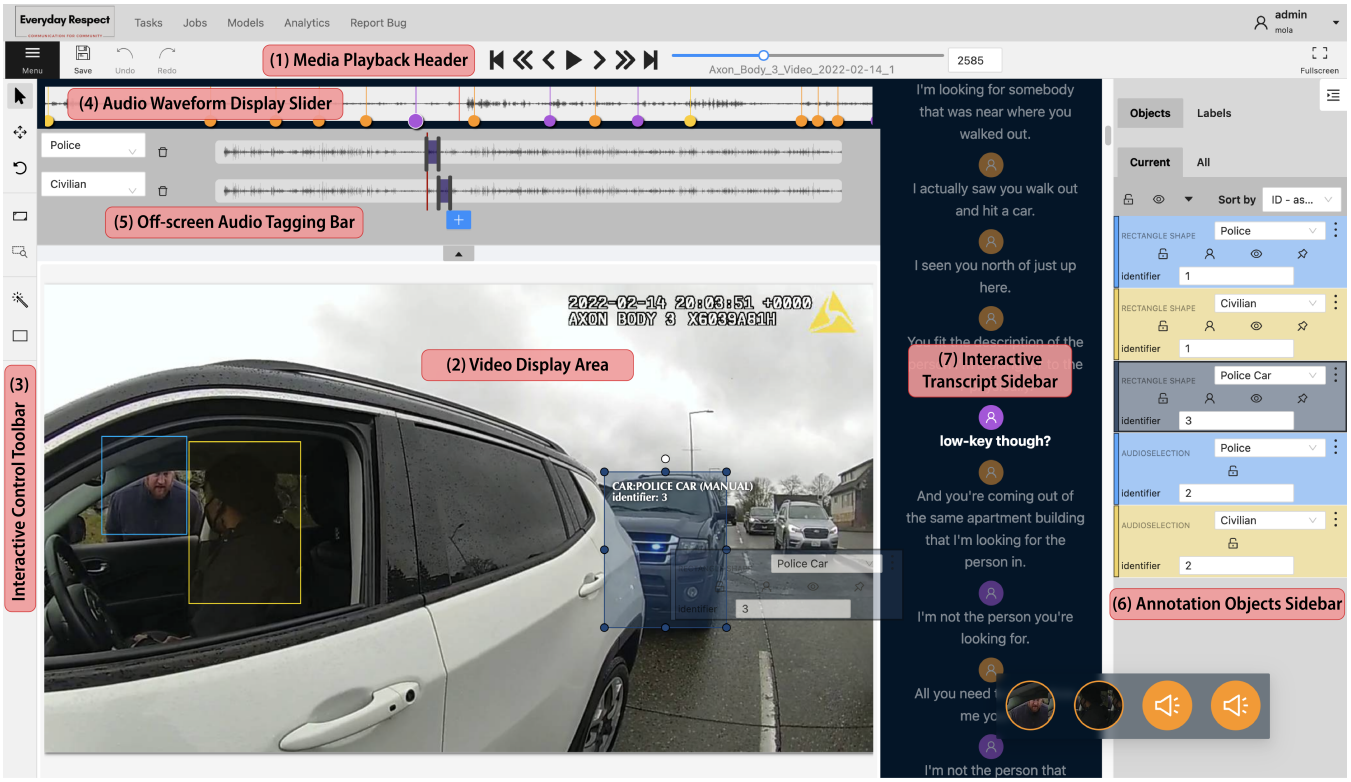

Figure 1: Layout of the User Interface, illustrating various components: (1) Media Playback Header, (2) Video Display Area, (3) Interactive Control Toolbar, (4) Audio Waveform Display Slider, (5) Off-screen Audio Tagging Bar, (6) Annotation Objects Sidebar, and (7) Interactive Transcript Sidebar. The screenshot features a video cleared for public release by the Seattle Police Department.

2 Platform Description

CVAT-BWV[2] is an open-source, web-based application that can be hosted on a local network, eliminating the need for an internet connection and thereby ensuring data privacy. Additionally, it incorporates user management to control access to annotator accounts and their assigned tasks. As a multimodal platform, it supports annotations for transcripts, auditory content, and visual entities, all of which are enhanced by AI-assisted annotation tools.

2.1 AI-Assisted Annotation Tools

Incorporating Human-in-the-Loop components in annotation tasks can enhance data quality and user experience while reducing the annotators' workload [van der Wal *et al.*, 2021; Marchesoni-Acland and Facciolo, 2023; Weber and Plank, 2023]. CVAT-BWT employs several AI models to assist with various annotation tasks:

ASR and Speaker Diarization. The platform leverages the audio channel in BWV footage to transcribe and segment the audio stream based on speaker identity. This is achieved through the `WhisperX` model [Bain *et al.*, 2023], which provides accurate word-level timestamps.

Object Detection. For automated bounding box generation around people and vehicles, the platform integrates `YOLOv8`, a state-of-the-art object detection model [Jocher *et al.*, 2023].

[2]https://github.com/USC-CSSL/CVAT-BWV

Face Recognition. To avoid duplicate bounding boxes for the same individual, newly created bounding boxes of people are fed into the `LightFace` face recognition framework [Serengil and Ozpinar, 2020] to be matched with previously created bounding boxes. We use the `RetinaFace` [Deng *et al.*, 2020] together with the `FaceNet` face recognition model [Schroff *et al.*, 2015] from the framework.

2.2 User Interface

The main annotation page of the tool comprises seven principal components (see Figure 1). ***(1) Media Playback Header*** offers controls for playing the video at regular speed or frame by frame. Annotators can navigate to a specific frame using the frame selector, control buttons, or the playback slider. Additionally, this header includes options for accessing the menu, saving progress, and performing undo or redo actions. BWV footage is displayed in the ***(2) Video Display Area***, where annotators can create and modify visual annotation objects (bounding boxes). This canvas also presents the bounding boxes overlaid on each frame of the video, complete with their assigned labels and identifiers. ***(3) Interactive Control Toolbar*** provides controls for annotators to drag, zoom, and rotate the video canvas. It also provides tools to manually add bounding boxes by drawing over the Video Display Area or through an automated object detection tool. ***(4) Audio Waveform Display Slider*** illustrates the audio waveform's amplitude variations over time in the video. It also displays

Figure 2: Illustration of the CVAT-BWV annotation workflow. The process begins with uploading BWV footage and proceeds through ASR & Speaker Diarization, annotation of transcripts, creation of bounding boxes and audio tags, linking transcripts to visual and auditory annotations, and comes to an end with the completion of targeted questionnaires for each annotation object.

transcript segments with color-coded handles—termed "transcript timing handles." Annotators can create a new transcript or adjust the start and end times of an existing transcript by dragging these handles along the waveform. ***(5) Off-screen Audio Tagging Bar*** is used when there is an audible utterance in the video whose speaker is not visible. In such instances, annotators can create auditory annotation objects, known as audio tags. Using the "audio timing handles" on the waveform track, annotators can adjust the start and end times of the utterance and assign labels to the audio tags. ***(6) Annotation Objects Sidebar*** lists bounding boxes and audio tags created by annotators for the current frame or throughout the entire video. Each entry on this list corresponds to an annotation object and includes a button for navigating to the precise frame of its creation. The assigned label or identifier for each entry can be modified, or the entry can be deleted if necessary. ***(7) Interactive Transcript Sidebar*** displays the complete transcript of the video, highlighting the currently spoken utterance. Initially, this sidebar is populated with the output from the ASR module, providing an auto-generated transcript and the results of speaker diarization, each marked with a unique colorful avatar. Annotators can then correct any errors. Moreover, inaccuracies in speaker diarization can be rectified by reassigning colors to utterances.

2.3 Workflow

Task Creation and Assignment. In CVAT-BWV, annotation begins with the task creation step, where raw BWV footage is uploaded, label categories and subcategories can be modified or created, and tasks can be assigned to annotators. While the video is uploading, the ASR and speaker diarization module extracts transcripts and identifies speakers in the background.

Transcript Annotation. After task assignment, annotators can access the task. The *Interactive Transcript Sidebar* displays utterances from detected speakers, each represented by a uniquely colored avatar. Additionally, transcript timing handles appear in the *Audio Waveform Display Slider*, color-coded to match the avatars of the speakers.

Creation of Bounding Boxes and Audio Tags. On each video frame, annotators have the option to pause and create visual or auditory annotation objects. Bounding boxes can be manually drawn over a frame or automatically generated using the object detection module for identifying people and vehicles. Audio tags are created using the *Off-screen Audio Tagging Bar* and the audio timing handles on the waveform tracks. Annotators then assign one of the predefined labels to each annotation object created during the task creation step. If a "Person" tag is assigned to a bounding box, the face recognition module processes it to check for duplicates among previously created face bounding boxes, alerting the annotator with a warning in case of a previously tagged person. However, the decision to dismiss the warning is left to the annotator, given the face recognition module's potential inaccuracies. The de-duplication process is repeated upon task completion to ensure no multiple bounding boxes have been created for a single individual.

Linking Transcripts to Annotation Objects. At any point, it is possible to link a uniquely colored avatar from the transcripts sidebar to one of the visual or auditory annotation objects, with this assignment automatically applied to all avatars of the same color.

Annotation Object Questionnaires. After the creation of annotation objects, annotators can be prompted to complete questionnaires specific to each created annotation object, whether a bounding box or an audio tag, based on their assigned label.

3 Conclusion

Many of the most consequential government-civilian interactions in our democracy, including those involving police, are captured on body-worn cameras or similar devices. The data from these devices hold enormous potential to facilitate transparency and accountability efforts, but only if effective ML tools can be developed to analyze this footage. Effective tools to analyze these complex social interactions begin with extensive human annotation.

The CVAT-BWV platform introduced here provides four core features not available in current annotation tools: 1) The ability to operate in secure data environments suitable for sensitive or classified data; 2) Flexible customization and AI integration for automatic labeling of domain-relevant audio and video features; and 3) Support for multimodal and multi-perspective annotations. Our tool is released under an MIT/X license, and we hope its widespread adoption will facilitate rapid progress in developing the tools necessary to empower the use of BWVs as a tool for learning and oversight.

Ethical Statement

The nature of the BWV footage, often capturing highly sensitive and potentially violent interactions, necessitates careful consideration of the annotators' well-being. Acknowledging the emotional and psychological toll that extended exposure to such content can have, it is crucial to monitor for signs of secondary trauma or post-traumatic stress disorder among those tasked with annotation. Implementing measures to support mental health and mitigate adverse effects is not only ethical but essential for responsibly handling the annotation of BWV data.

Acknowledgments

We thank our anonymous reviewers for their valuable feedback. This work was funded by the Microsoft Justice Reform Initiative, National Science Foundation Civic Program, Arnold Ventures, Google Award for Inclusion Research, and USC Zumberge Interdisciplinary Grant. The views expressed are those of the authors and do not necessarily represent the positions of the funders.

Contribution Statement

Authors Preni Golazizian, Rajat Hebbar, and Jackson Trager contributed equally to this work.

References

[Bain *et al.*, 2023] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Proc. INTERSPEECH 2023*, pages 4489–4493, 2023.

[California Department of Justice, 2024] California Department of Justice. Racial and identity profiling advisory board annual report. https://oag.ca.gov/system/files/media/ripa-board-report-2024.pdf, 2024. Accessed: 2024-02-05.

[Camp *et al.*, 2021] Nicholas P. Camp, Rob Voigt, Dan Jurafsky, and Jennifer L. Eberhardt. The thin blue waveform: Racial disparities in officer prosody undermine institutional trust in the police. *Journal of Personality and Social Psychology*, 121(6):1157–1171, December 2021.

[Deng *et al.*, 2020] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[Graham *et al.*, 2024] Benjamin AT Graham, Lauren Brown, Georgios Chochlakis, Morteza Dehghani, Raquel Delerme, Brittany Friedman, Ellie Graeden, Preni Golazizian, Rajat Hebbar, Parsa Hejabi, et al. A multi-perspective machine learning approach to evaluate police-driver interaction in Los Angeles. *arXiv preprint arXiv:2402.01703*, 2024.

[Jocher *et al.*, 2023] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023.

[Labelbox, 2024] Labelbox. Labelbox. https://labelbox.com, 2024. [Online; accessed 2024].

[Marchesoni-Acland and Facciolo, 2023] Franco Marchesoni-Acland and Gabriele Facciolo. Iadet: Simplest human-in-the-loop object detection, 2023.

[Montani and Honnibal, 2024] Ines Montani and Matthew Honnibal. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models, 2024.

[Prabhakaran *et al.*, 2018] Vinodkumar Prabhakaran, Camilla Griffiths, Hang Su, Prateek Verma, Nelson Morgan, Jennifer L Eberhardt, and Dan Jurafsky. Detecting institutional dialog acts in police traffic stops. *Transactions of the Association for Computational Linguistics*, 6:467–481, 2018.

[Rho *et al.*, 2023] Eugenia H. Rho, Maggie Harrington, Yuyang Zhong, Reid Pryzant, Nicholas P. Camp, Dan Jurafsky, and Jennifer L. Eberhardt. Escalated police stops of black men are linguistically and psychologically distinct in their earliest moments. *Proceedings of the National Academy of Sciences*, 120(23):e2216162120, 2023.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[Sekachev *et al.*, 2020] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, August 2020.

[Serengil and Ozpinar, 2020] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.

[van der Wal *et al.*, 2021] Douwe van der Wal, Iny Jhun, Israa Laklouk, Jeff Nirschl, Lara Richer, Rebecca Rojansky, Talent Theparee, Joshua Wheeler, Jörg Sander, Felix Feng, Osama Mohamad, Silvio Savarese, Richard Socher, and Andre Esteva. Biological data annotation via a human-augmenting AI-based labeling system. *NPJ Digit. Med.*, 4(1):145, October 2021.

[Voigt *et al.*, 2017] Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526, 2017.

[Weber and Plank, 2023] Leon Weber and Barbara Plank. ActiveAED: A human in the loop improves annotation error detection. In Anna Rogers, Jordan Boyd-Graber, and

Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8834–8845, Toronto, Canada, July 2023. Association for Computational Linguistics.