Towards Smarter Data Handling in Law Enforcement: Temporal Action Localization for Body-Worn Cameras

Spencer Lee
Department of Electrical and Computer Engineering
Faculty of Applied Science
University of British Columbia
Vancouver, Canada
slee67@student.ubc.ca

***Abstract*—Body-Worn Cameras (BWCs) have become indispensable tools for law enforcement agencies, capturing crucial evidence and enhancing transparency. However, the proliferation of BWCs has generated vast amounts of video data, posing significant storage and analysis challenges. In this paper, we present an approach to address these challenges through Temporal Action Localization (TAL), a machine learning technique that detects and timestamps specific actions or events within BWC footage. We outline the methodology, encompassing dataset collection, architecture selection, and annotation techniques, all tailored to the unique demands of BWC data. Our research highlights the importance of adaptability and generalization in selecting machine learning architectures and discusses the challenges of annotating diverse law enforcement scenarios. Furthermore, we offer insights into future work directions, such as customized batch loading, object classification, and dataset generalization, to advance the field of BWC data management and analysis. By leveraging TAL and innovative strategies, we aim to optimize the utility of BWCs in enhancing public safety, accountability, and law enforcement practices.**

***Index Terms*—machine learning, temporal action localization, computer vision, body-worn cameras, dataset collection, annotation, video analysis, data management**

I. Introduction

Recent years have witnessed the widespread adoption of body-worn cameras (BWCs) within police forces across the world, with a particular emphasis on their use in the United States. Key providers, such as Axon, have played a pivotal role in supplying both BWCs and cloud-based storage solutions to more than 14,000 police agencies. The cumulative result of this extensive deployment is the generation of an astounding 120,000 terabytes (TB) of video data [1]. This data influx poses a formidable challenge: how can this vast volume of video data be efficiently filtered or compressed to prioritize critical footage while deprioritizing redundant or unimportant content?

One promising avenue for addressing this challenge lies in machine learning (ML) approaches; specifically, the application of Temporal Action Localization (TAL) techniques to BWC footage. TAL is a machine learning objective that seeks to answer the fundamental questions of "when" an action takes place and "what" category of action is occurring [2]. Traditionally, TAL has been employed to detect the start and end of human movements, such as pitching a baseball, and for scene segmentation in videos.

It has, however, seen minimal use in practical applications due to its limitations. A significant hurdle in adopting TAL techniques for BWC data management is the prevalent closed-world assumption. In this context, TAL models are developed with the presupposition that all action categories are known a priori. Real-world scenarios, especially in law enforcement, often involve open-world settings where novel action categories can emerge without prior training data. This raises a crucial question: can existing TAL techniques be adapted to semi-closed-world scenarios, such as those encountered in policing, to benefit commercial goals?

The overarching problem addressed in this research is that police agencies are in urgent need of a solution to alleviate the storage strain on their digital infrastructure by reducing redundant or unimportant body-camera video. As the adoption of BWCs continues to grow, this challenge has become increasingly pressing, with both large and small police agencies grappling with the storage dilemma [3]. Large agencies, due to the sheer number of active officers, generate a constant stream of BWC footage, while smaller agencies may lack the necessary infrastructure for efficient storage [4]. Moreover, legal regulations mandate the retention of BWC footage for specified durations, precluding the simple deletion of older data when storage space runs low. The use of proprietary codecs by some BWC manufacturers further complicates data compatibility issues, highlighting the need for a standardized video format.

This paper endeavours to explore a solution to the problem at hand by employing machine learning-based TAL techniques. The core objective is to investigate whether TAL models can effectively identify and prioritize relevant actions within BWC footage, thus reducing the storage strain on police digital infrastructure. Additionally, this research aims to address the adaptability of TAL techniques to semi-closed-world scenarios, where novel action categories may emerge without prior

knowledge. Through these investigations, this paper seeks to provide valuable insights and practical solutions to the challenges faced by police agencies in managing BWC data effectively.

In the following sections, we will delve deeper into the background of police body-worn cameras, machine learning approaches applicable to this context, and the methodology employed in this research to address the identified problem. The subsequent chapters will further expound on the research findings, implications, and potential future directions, ultimately contributing to the discourse on BWC data management and machine learning in law enforcement.

Problem statement: *Police agencies need a way to alleviate the storage strain on their digital infrastructure by reducing redundant or unimportant body-camera video.*

II. Background

A. Police Body-Worn Cameras

The widespread adoption of body-worn cameras (BWCs) has gained momentum on a global scale, with a primary focus on their utilization within North American law enforcement agencies. These devices have been strategically deployed with two primary objectives: enhancing the safety of police officers and ensuring greater accountability for police interactions with the general public.

Among the prominent providers of BWCs, Axon has emerged as a market leader, offering products like the Axon Body 2 product line [5]. This dominance is reflected in their widespread adoption by a multitude of police agencies. Consequently, the extensive usage of BWCs translates into a staggering volume of video data, with thousands of terabytes being generated from these devices.

However, the challenge of managing this voluminous data is not mitigated by the size of the police agency. Large law enforcement organizations generate substantial amounts of BWC footage due to a higher number of active officers, while smaller agencies, although dealing with less data, often lack the robust infrastructure required for efficient data storage and management [4].

Moreover, legislative and policy requirements mandate the retention of BWC footage for specific timeframes. For example, in British Columbia, Canada, any BWC video must be retained for one year from the day after it was recorded [6]. In the case of the New York Police Department (NYPD), video recordings must be retained for a minimum of 18 months, with certain categories of footage being preserved for even longer durations [7]. This legal obligation prevents the simple deletion of older video data when storage space becomes scarce.

Another complicating aspect in BWC data management arises from the utilization of proprietary codecs and containers by certain BWC manufacturers [8]. This introduces compatibility challenges, underscoring the significance of adopting a standardized video format to enhance the efficiency of data handling. Furthermore, within the context of this project, these compatibility concerns play a pivotal role in shaping the format-handling capabilities of the BWC data management system.

Additionally, BWC recordings often contain audio that may be unclear, indecipherable, or even nonexistent, necessitating a data analysis approach that relies solely on video footage.

In response to these challenges, various technologies and solutions have been adopted within law enforcement agencies. AXON, for instance, provides cloud storage services to facilitate the storage and management of BWC footage [9]. Large technology corporations such as Microsoft and Google also host cloud servers specifically designed for storing police video data. Moreover, solutions like VIDIZMO [10] leverage artificial intelligence to search and index stored videos, making it easier for law enforcement agencies to access and analyze their recorded content.

It's worth noting that the approach to sharing BWC footage publicly varies across different jurisdictions. For example, the New York Police Department (NYPD) publishes heavily edited versions of BWC footage, often accompanied by commentary from the chief of police [11]. Similarly, the Los Angeles Police Department (LAPD) has started to release select videos, although the number of videos publicly available remains limited [12]. In contrast, Canada follows stringent privacy regulations, releasing BWC videos only in accordance with the country's Privacy Act and Personal Information Protection and Electronic Documents Act [13] [14]. Chicago's Civilian Office for Police Accountability (COPA), on the other hand, releases all videos in accordance with their Video Release Policy, with censorship redactions made to protect personal privacy [15].

Additionally, many jurisdictions limit the public posting of BWC footage in cases where no incident occurs. Some policies mandate that BWC recordings should only begin under specific circumstances, such as during mental health calls, interactions with individuals in crisis, crimes in progress, investigations, public disorder events, and protests, among others. These policies emphasize the importance of recording information to support law enforcement duties, capture interactions with individuals in custody, and document situations where audio and/or video evidence may be crucial for lawful execution of duties.

B. Machine Learning Approaches

In the context of managing police BWC footage, machine learning approaches play a pivotal role, primarily falling under the overarching domain of Computer Vision with a focus on video analysis. To address the challenge of storage space requirements in BWC data, various machine learning tasks come into play, each contributing uniquely to the broader goal of optimizing data management.

Temporal Action Localization (TAL): TAL is a key machine learning task that aims to detect activities within a video stream and provide precise timestamps indicating when these activities begin and end. In the context of police BWC footage, TAL can help identify specific actions or events of interest, such as arrests, interactions with the public, or critical incidents. By pinpointing these moments in the video,

it becomes possible to extract and prioritize relevant segments, thus reducing the storage strain caused by retaining redundant or unimportant footage. TAL allows for efficient data retrieval when needed, contributing to streamlined data management.

Video Understanding: Video Understanding, as a machine learning task, goes beyond mere action localization. It seeks to recognize and spatially-temporally localize various actions or events within a video. In the context of BWC data management, this task can be employed to comprehensively understand the content of recorded videos. This understanding facilitates the automatic categorization of video segments based on the actions or events they contain. By categorizing video segments, the system can make informed decisions about which footage to prioritize for storage or retrieval, ensuring that critical incidents are retained while non-essential content is deprioritized or compressed.

Action Classification: Action Classification is another crucial machine learning task with the objective of categorizing videos based on the actions or events they portray. In the context of police BWC data, this task plays a pivotal role in classifying recorded interactions or incidents into predefined action categories. This categorization empowers the system to discern and give priority to videos aligned with specific categories of interest, such as ”arrests,” ”traffic stops,” or ”public protests.” Through the classification of BWC footage, the data management system can efficiently structure and archive videos according to their relevance to distinct policing scenarios, thus enhancing storage optimization. It’s noteworthy that action classification primarily operates at the video level for recognition, implying that it classifies entire BWC recordings rather than individual video segments within them.

Action Recognition: Action Recognition extends the capabilities of machine learning to recognize human actions within videos or images and classify them into predefined action classes. In the context of police BWC footage, this task aids in understanding the behaviours and actions of individuals involved in recorded incidents. It enables the system to automatically identify actions such as ”compliance,” ”confrontation,” or ”use of force.” Action Recognition enhances the granularity of data analysis, allowing for fine-grained categorization and prioritization of video segments. This level of detail ensures that critical actions are captured and retained while irrelevant or redundant content is efficiently managed.

Furthermore, in addition to incorporating relevant machine learning tasks, it is essential to utilize a powerful technique known as Parallel Multi-Head Attention. This technique, widely applied in machine learning, especially in the domains of natural language processing and computer vision, greatly enhances a model’s capacity to simultaneously focus on different aspects of the input data. By employing multiple attention heads, each assigned to learn distinct features, the model achieves a more comprehensive grasp of the data’s intricacies. In the context of police BWC data management, parallel multi-head attention can be leveraged to concurrently process various facets of the video, including temporal features, spatial information, and cues for action recognition. This concurrent processing has the potential to significantly enhance the accuracy of critical tasks such as action recognition and localization.

Additionally, corollary datasets play a crucial role in identifying optimal pre-existing machine learning architectures for adaptation to the specific requirements of BWC footage analysis. These datasets serve as valuable indicators of the most effective machine learning models within the predefined tasks of temporal action localization, video understanding, action classification, and action recognition. By evaluating the performance of various architectures on these corollary datasets, we can discern which models excel in tasks closely related to BWC footage analysis. This insight guides the selection of the most suitable machine learning architecture, which can then be tailored to address the unique challenges and demands posed by BWC data management and analysis.

In summary, machine learning tasks in Computer Vision, combined with techniques like parallel multi-head attention and reference to top corollary datasets, contribute significantly to the overarching goal of optimizing police BWC data management. These tasks enable the system to intelligently process, categorize, and prioritize video segments, reducing the storage strain while ensuring the retention of critical incident footage.

C. Annotation Tools

Effective annotation tools are a essential to any machine learning project, especially when dealing with video data as opposed to images. In the context of annotating BWC footage, there are specific requirements and considerations to ensure the accuracy and efficiency of the annotation process.

Temporal annotations are a fundamental requirement for annotation tools in the context of managing BWC data. Since BWC footage comprises video clips, it is imperative for annotators to delineate specific time intervals within the videos, indicating precisely when particular actions or events transpire. These temporal annotations serve as the essential reference points for training machine learning models, enabling them to accurately identify and categorize actions inferred from their ground truth training. This form of annotation proves especially crucial when the goal is to perform video analysis at a level of granularity finer than the video as a whole. In the context of this project, the ability to adjust the bitrate within a video hinges on the presence of precise temporal annotations, underscoring their significance in facilitating more nuanced data management and analysis.

Object Classification becomes necessary in certain scenarios, where annotators must be able to temporally and spatially classify objects within video frames. This involves the recognition and labelling of objects of interest, which may include vehicles, weapons, or specific items pertinent to law enforcement contexts. An annotation tool equipped for object classification should possess the capability to draw bounding boxes around these objects within frames and track them as they persist across successive frames. This functionality ensures that the annotation tool can efficiently capture both

temporal and object-specific information in a synchronized manner, facilitating a more comprehensive understanding of the video content.

In addition to annotation tool functionalities, compatibility with the Windows 10 operating system is imperative. To ensure seamless integration with the existing technology infrastructure, the chosen annotation tool should be compatible with the Windows 10 operating system. This compatibility ensures that the annotation process can be efficiently carried out on the preferred operating system used by many law enforcement agencies.

Several publicly available annotation tools exist, and among them is the ViPER-GT annotation tool (ViPER) [16] [17], which provides a Java-based graphical user interface for authoring ground truth data. This tool is specifically designed for the frame-by-frame markup of video metadata stored in ViPER's custom format, making it valuable for tasks that require detailed video annotation. However, it's important to note that ViPER is considered somewhat outdated when compared to more modern annotation tools. Its limited feature set and lack of customization options may limit its suitability for the complex annotation requirements associated with BWC footage. Nonetheless, it can still serve as a useful tool for basic video metadata visualization and annotation tasks.

FiftyOne - also known as Voxel51 - is a Python library designed to facilitate data exploration, visualization, and annotation for machine learning datasets [18]. It offers a range of features that align with the needs of BWC data management:

Customizability: FiftyOne provides customizable solutions for annotating video data, making it well-suited for the unique requirements of BWC footage. Annotators can define specific annotation types, including temporal annotations and object classification tags, to capture the relevant information within the videos.

Integration with CVAT: FiftyOne integrates seamlessly with the Computer Vision Annotation Tool (CVAT), a powerful open-source annotation platform. This integration enhances the annotation process by providing access to CVAT's robust annotation capabilities while leveraging FiftyOne's data exploration and visualization functionalities.

Python-Based: Being a Python library, FiftyOne aligns with the programming language commonly used in machine learning and computer vision tasks. This ensures that annotation processes can be streamlined and integrated into the broader machine learning pipeline.

Custom Dataset Support: FiftyOne allows for the creation of custom datasets, enabling researchers and annotators to structure and manage video annotations efficiently. This custom dataset support is crucial when handling the diverse and extensive BWC video data. It also is crucial when defining the output format of the annotations so that they can be easily interpreted by specific machine learning architectures.

In summary, annotation tools are essential components of the BWC data management process, facilitating the creation of accurate and informative ground truth annotations. While older tools like ViPER may fall short of meeting the complex requirements of BWC data, modern solutions like FiftyOne offer the customizability and integration needed to efficiently annotate and prepare this unique dataset for machine learning tasks.

III. Methodology

A. Dataset Collection

The process of dataset collection is a crucial step in this research, as the lack of an existing dataset necessitates sourcing, downloading, and annotating a new set of videos. To create a comprehensive dataset for this study, the Chicago Body-Worn Camera (BWC) footage was selected due to its diversity in scenarios and a consistent level of clip quality. All videos in this dataset were recorded using the same resolutions and overlay information, ensuring uniformity in video characteristics. However, certain videos within this dataset vary in their video formats due to the change and upgrade body camera models throughout the years.

To efficiently acquire the Chicago BWC footage, the *yt-dlp* tool [19] was employed as a versatile web scraping and video downloading solution. To ensure relevance and accuracy in video selection, *yt-dlp* was configured to download videos specifically tagged with body cam descriptions. This filtration process was essential, as the dataset contains a mix of surveillance videos, which were not pertinent to the research objectives, and therefore needed to be filtered out.

However, it's worth noting that the Chicago BWC database presented certain access challenges. The videos were hosted on Vimeo, and access to the database required a Vimeo account. To circumvent this limitation, authentication credentials were integrated into the *yt-dlp* program, either directly or via browser cookies. This allowed for seamless access to the video content.

One significant hurdle in dataset preparation was the variation in video encoding formats across the Chicago BWC footage. To ensure consistency and compatibility within the dataset, all videos were re-encoded to a standardized format. MP4 was chosen for its widespread compatibility and ease of use. This re-encoding process was facilitated using the FFMPEG tool, ensuring that all videos within the dataset adhered to a common format.

By meticulously addressing these challenges and following a systematic approach to dataset collection, a high-quality and standardized dataset was created for subsequent stages of research, including annotation and machine learning model development. This dataset serves as the foundation for training and evaluating machine learning algorithms for the efficient management of BWC data.

The curated Chicago BWC dataset, comprising a diverse range of video clips, was securely stored and managed within the cloud-based infrastructure provided by the Digital Research Alliance of Canada [20]. Leveraging the robust capabilities of cloud computing, the dataset was organized, catalogued, and made readily accessible for research purposes. This cloud-based storage solution ensured the preservation and accessibility of the dataset, regardless of geographical

locations. Moreover, cloud storage offered the advantage of scalability, enabling the dataset to be expanded as needed to accommodate future research requirements. The use of cloud-based storage within a research environment underscores the importance of modern infrastructure in supporting data-intensive projects while ensuring data security and accessibility.

B. Machine Learning

Machine learning (ML) played a pivotal role in this research, specifically in addressing the challenge of managing and reducing the storage strain caused by the massive influx of BWC footage. The ML approach was prioritized after the comprehensive collection and curation of the BWC dataset. One of the key considerations in this phase was the selection of an appropriate ML architecture. To ensure maximum flexibility and compatibility, the choice of architecture was only limited by the format of the video data, which had been standardized to MP4 during the dataset preprocessing stage.

The selection of the ML architecture was a critical decision, as it would significantly impact the format and structure of the annotations that would be used to train and evaluate the model. Annotations, such as temporal action localization (TAL) and action classification, are essential for teaching the model to recognize and categorize actions accurately within the BWC footage. Given the importance of annotations, the ML architecture was chosen before the development of the annotation platform.

In the pursuit of an effective ML architecture, we adopted a strategy of repurposing pre-existing architectures that had demonstrated excellence in various video computer vision tasks. Rather than reinventing the wheel, we sought architectures that had excelled in related domains, such as temporal action localization (TAL), video understanding, action classification, and action recognition.

Notably, while video understanding is a crucial aspect of BWC data management, it lacks well-known benchmark datasets that would have facilitated the selection of a dedicated architecture. To overcome this limitation, we leveraged related video subtasks, particularly TAL, action classification, and action recognition, to guide our architecture selection process. These subtasks offered established benchmarks and datasets against which we could evaluate the performance of various architectures.

Within these three primary categories, several benchmark datasets were available, each designed to represent specific aspects of video structures and actions. For instance, THUMOS'14 featured human actions in diverse environmental settings, comprising a relatively low number of actions (approximately 20) with multiple videos depicting each action. The dataset included temporal action annotations, which detailed the start and end times of each action within the videos.

In contrast, the Kinetics-700 dataset encompassed a large number of actions (around 700) exhibited across various datasets. However, this dataset primarily provided annotations at the video level, associating each video with a specific action class that appeared within its footage.

To establish a basis for selecting the most suitable ML architecture, we meticulously considered benchmark datasets that represented distinct video subtasks. From each of the three primary categories—TAL, action classification, and action recognition—we selected the two most popular benchmarks renowned for their rigorous evaluation criteria and large-scale datasets. These benchmarks provided a diverse range of video structures and actions, ensuring that the chosen ML architecture would be versatile enough to excel in various BWC data analysis tasks.

The devised weighted matrix shown in Table I was instrumental in our selection process. For each architecture under consideration, we calculated a weighted score based on its mean Average Precision (mAP) performance in relation to the top-scoring mAP within the corresponding benchmark. This approach allowed us to objectively compare architectures across different benchmarks. In the TAL category, we examined THUMOS'14 [21], a benchmark characterized by its collection of human actions performed in diverse environmental settings. THUMOS'14 featured a relatively modest number of actions (approximately 20) but included multiple videos illustrating each action. Simultaneously, ActivityNet-1.3 [22], another benchmark within the TAL category, challenged the ML architectures with its extensive dataset of temporal action annotations.

Within the realm of action classification, we evaluated the Kinetics-400 [23] benchmark, which comprised a substantial number of action classes. Furthermore, we explored Charades [24], a benchmark that introduced complexities by featuring actions embedded in real-life video scenes, representing an environment akin to BWC footage.

Finally, for action recognition, we considered Something-Something V2 [25] and AVA v2.2 (Atomic Visual Actions) [26]. These benchmarks were selected for their emphasis on recognizing and categorizing a wide array of actions, often with complex spatial and temporal dependencies.

It's important to emphasize that while these benchmarks were instrumental in shaping our architecture selection, some architectures were not represented in all of these benchmarks. To underscore the importance of an architecture's generalization capabilities, we introduced a novel approach within the weighted matrix. In cases where a benchmark was missing a certain ML architecture, we adopted a fair approach by not directly imposing penalties on these architectures' weighted scores. Instead we incorporated an additional column that ranked ML architectures based on the number of missing benchmarks, from the least to the most. This innovative column served as a decisive factor in our selection process, further highlighting the significance of an architecture's adaptability and versatility. By considering not only the performance within benchmarks but also the ability to excel across a diverse range of video subtasks, we ensured that the chosen ML architecture would be exceptionally well-suited to tackle the multifaceted challenges of BWC data analysis.

TABLE I
MACHINE LEARNING ARCHITECTURE WEIGHTED MATRIX BASED ON BENCHMARK SCORES

	Temporal Action Localization		Action Classification		Action Recognition			
	THUMOS'14	ActivityNet-1.3	Kinetics-400	Charades	Something-Something V2	AVA v2.2	Weighted Score	Missing Benchmarks
InternVideo	1.000	0.929	1.000		0.999	0.909	0.967	1
VideoMAE	0.972		0.988		0.996	0.945	0.975	2
I3D	0.933	0.869	0.853	0.757			0.853	2
TubeVIT			0.998	0.998	0.984		0.994	3
Hiera			0.964		0.990	0.960	0.971	3
MVD			0.957		1.000	0.911	0.956	3
MoViNet			0.895	0.953	0.821		0.890	3
TokenLearner			0.937	1.000			0.969	4
UMT			0.995			0.882	0.938	4
TriDet	0.979	0.876					0.928	4
PRN		1.000					1.000	5
LART						1.000	1.000	5
TCANet		0.894					0.894	5

C. Annotation

Once the videos were downloaded and ML architecture chosen, the crucial step of annotation became necessary to provide the ground truth for training and evaluating machine learning models. Initially, the ViPER annotation tool was employed for this purpose. However, it quickly became apparent that ViPER's outdated interface and lack of customizability in output formats posed significant limitations.

To address these limitations, an alternative annotation solution was sought, leading to the adoption of the FiftyOne python library. FiftyOne not only offered the flexibility and customization capabilities that were lacking in ViPER but also had the advantage of being regularly updated to align with the latest Python releases. This made it a more suitable choice for the dynamic requirements of the project.

To utilize FiftyOne effectively, a connection to the Computer Vision Annotation Tool (CVAT) was necessary. Access to CVAT servers was granted through a free account, which allowed for the seamless integration of CVAT into the annotation workflow. However, it's worth noting that the use of CVAT servers came with certain limitations, primarily regarding the number of annotations that could be performed.

To circumvent these limitations and ensure the comprehensive annotation of the dataset, a local server running CVAT was set up. This was achieved by deploying CVAT on a Dockered Windows Subsystem for Linux (WSL) server within the Ubuntu operating system environment. This local server configuration allowed for more extensive annotation work without restrictions on the number of annotations, effectively removing a significant obstacle.

In addition to leveraging FiftyOne for annotation, a custom annotation interface was developed to streamline the annotation process further. This interface was designed in collaboration with FiftyOne library functions to provide a tailored and efficient environment for annotators. Furthermore, custom format exporter and saver classes were created to handle the video annotations correctly. These classes included a custom sorting method for dataset organization based on video size and the number of annotations. Additionally, a custom annotation script was implemented to streamline the annotation of videos without existing annotations and to allow for the modification of pre-existing annotations when needed. The project also introduced a custom dataset class, building upon FiftyOne's *TemporalDetection* class, designed to accommodate the unique demands of the project. Lastly, a custom dataset exporter was developed, capable of handling the custom dataset and generating annotations in a text file format customized to align with the project's specific annotation needs.

IV. CONCLUSION

In this study, we embarked on a journey to explore the potential of machine learning (ML) approaches in mitigating the storage strain associated with Body-Worn Camera (BWC) footage, a challenge faced by law enforcement agencies worldwide. Our comprehensive methodology encompassed various critical stages, from dataset collection to ML architecture selection, and finally, the intricate process of annotation. Throughout this journey, we encountered both triumphs and challenges, shedding light on the complexities of adapting ML to the unique demands of BWC data.

One of the pivotal milestones in our research was the selection of the InternVideo ML architecture ([27], [28]), a decision guided by the insightful weighted matrix. InternVideo's impressive performance within benchmark datasets, coupled with

its excellent documentation and annotation format examples, made it a compelling choice. However, our journey took a nuanced turn when we discovered that adapting InternVideo to BWC footage was not without its hurdles.

The custom dataset class designed within InternVideo's helper functions marked a significant step toward BWC data integration. Yet, we encountered a obstacle: the sheer length of BWC footage sequences. These extended video segments, often capturing complex law enforcement interactions such as arrests spanning minutes to half an hour, presented a formidable challenge. Attempting to process and store these lengthy videos within the constraints of available memory led to out-of-memory errors, creating a roadblock in our efforts.

The crux of our findings underscores a fundamental limitation when adapting ML approaches to alleviate BWC storage strain. The necessity of extensive pre-processing and the handling of lengthy video data fundamentally contradicts the notion of ML automation. While ML excels in recognizing patterns, categorizing actions, and making sense of complex visual data, the burdensome preprocessing requirements inherent to BWC data negate the efficiency gains that automation can offer.

In addition to the findings and challenges outlined above, it's essential to acknowledge a significant accomplishment achieved during this research journey—the creation of a comprehensive Body-Worn Camera (BWC) dataset. This dataset, sourced from the Chicago Police Department, underwent a multifaceted annotation process that allowed us to gain valuable insights into the complexities of law enforcement scenarios.

Initially, our annotation efforts aimed to mark crucial events within the BWC footage, including arrests, shootings, and other high-danger situations. However, as we delved deeper into the task, we encountered several roadblocks. Interpreting these events accurately from a non-police perspective proved exceptionally challenging. Subjectivity loomed large when determining the onset and conclusion of these "important" events, and many of them spanned the entire duration of the video clips.

In light of these challenges, we pivoted our annotation strategy toward capturing the less conspicuous yet equally significant events occurring throughout the videos. Particular attention was given to scenes where law enforcement officers were in their patrol cars, as these segments contributed less to the semantic understanding of police incidents. This shift in focus allowed us to build a more rigorous dataset that was less prone to subjectivity.

In conclusion, this project has illuminated the intricate dance between ML's promise and the unique challenges posed by BWC data. While InternVideo's selection and generation of a novel BWC dataset marks a significant achievement, the relentless demands of lengthy BWC sequences underscore the paramount importance of developing innovative solutions that harmonize ML's capabilities with the practical realities of BWC data management.

V. Future Work

The path forward in the realm of Body-Worn Camera (BWC) data management and machine learning offers several promising avenues for future research and development.

Customized Batch Loader: One potential direction for future work involves the development of a customized batch loader within the InternVideo architecture. Currently, the architecture loads entire video clips into memory at once, posing challenges when handling long BWC footage sequences. By creating a batch loader that processes video clips in segments, this limitation can be addressed, allowing for more efficient and scalable data processing.

Object Classification: Expanding beyond temporal annotations, future research can delve into object classification within BWC footage. This entails identifying and classifying specific objects or entities within the video frames, such as vehicles, weapons, or other objects relevant to law enforcement scenarios. Developing robust object classification models can enhance the depth of understanding and analysis of BWC data.

Dataset Generalization: To further bolster the utility of BWC datasets, future work may involve appending the existing dataset with footage from different law enforcement jurisdictions. This expansion would contribute to dataset generalization, enabling the development of more versatile and adaptable machine learning models. Incorporating data from diverse sources can help address the challenges associated with varying recording conditions, equipment, and operational scenarios.

Audio Analysis: Integrating audio analysis into BWC data management and machine learning tasks can provide a more holistic understanding of law enforcement incidents. This includes recognizing spoken words, sounds of interest (e.g. gunshots, sirens), and sentiment analysis to gauge the emotional context of interactions.

Real-Time Analysis: Developing real-time analysis capabilities for BWC footage can enable law enforcement agencies to receive actionable insights during ongoing incidents. This involves the integration of machine learning models into BWC devices to provide immediate assistance and decision support to officers in the field.

Privacy Preservation: Future research can explore advanced techniques for preserving the privacy of individuals captured in BWC footage while still enabling effective analysis. This includes the development of anonymization methods and privacy-preserving machine learning algorithms.

These future work directions represent opportunities to advance the capabilities of BWC data management and machine learning, ultimately contributing to improved law enforcement practices, public safety, and accountability.

References

[1] Grant Fredericks. Video & policing 9/11 to today: The road to video literacy for investigators. *Police Chief*, 88, no. 9:102–106, 2021.

[2] Yaru Zhang, Xiao-Yu Zhang, and Haichao Shi. Ow-tal: Learning unknown human activities for open-world temporal action localization. *Pattern Recognition*, 133:109027, 2023.

[3] Lindsay Miller and Jessica Toliver. Implementing a body-worn camera program: Recommendations and lessons learned. *Police Executive Research Forum*, pages 32–33, 2014.

[4] Bill Hutchinson. Recent high-profile deaths put police body cameras under new scrutiny, Mar 2023.

[5] Inc. Axon Enterprise. Axon body 2. Product page on Website, 2023.

[6] Government of British Columbia. Provincial policing standards. Government Document, 2019.

[7] New York City Police Department. Body-worn cameras. Website, 2023.

[8] Muhammad Nabeel Ali. What challenges are faced when working with body camera video storage. Website, 2022.

[9] Inc. Axon Enterprise. Axon evidence. Product page on Website, 2023.

[10] VIDIZMO LLC. End-to-end evidence security & privacy. Website, 2023.

[11] NYPD. Nypd fid body-worn camera footage. Playlist on YouTube, 2023.

[12] Los Angeles Police Department. Critical incident community briefings. Playlist on YouTube, 2023.

[13] Government of Canada. Privacy act. Government Act, 1983. Accessed: October 5, 2023.

[14] Government of Canada. Personal information protection and electronic documents act. Government Act, 2000. Accessed: October 5, 2023.

[15] Civilian Office of Police Accountability. Case portal. Online Case Evidence Database, 2023. Accessed: October 5, 2023.

[16] David S. Doermann and David Mihalcik. Tools and techniques for video performance evaluation. *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, 4:167–170 vol.4, 2000.

[17] Vladimir Mariano, Junghye Min, Jin Park, Rangachar Kasturi, David Mihalcik, Huiping Li, David Doermann, and Thomas Drayer. Performance evaluation of object detection algorithms. In *Proceedings - International Conference on Pattern Recognition*, volume 3, pages 965–969, 01 2002.

[18] B. E. Moore and J. J. Corso. Fiftyone. *GitHub. Note: https://github.com/voxel51/fiftyone*, 2020.

[19] yt dlp. yt-dlp. GitHub Repository, 2023.

[20] Digital Research Alliance of Canada. Advanced research computing. Cloud Computing Platform, 2023.

[21] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2015.

[22] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[24] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ivan Laptev, Ali Farhadi, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *ArXiv e-prints*, 2016.

[25] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017.

[26] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions, 2018.

[27] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[28] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.