Real-Time Detection of School Bullying Using CCTV

Adams Ubini
University of Maryland, Baltimore County
ubiniadams@gmail.com

Yogesh Rawat
University of Central Florida
yogesh@crcv.ucf.edu

***Abstract*—Bullying in school institutions remains a persistent and often underreported issue, despite widespread CCTV deployment. Most school surveillance systems are passive and lack automated tools to detect aggressive behavior in real time. In this project, we explore a vision-language-based approach for detecting school bullying using CCTV footage. We curated a 12-class dataset from public datasets, movies, and social media, capturing a range of physical bullying actions alongside non-bullying interactions. Our dataset includes over 3,000 video clips labeled with fine-grained bullying actions such as "punching", "shoving", and "hair grab", as well as mixed-label behaviors. We fine-tuned a video action recognition model using 25 uniformly sampled frames per clip and trained for 20 epochs. The model achieved a baseline top-1 accuracy of 43%, with over 90% accuracy on non-bullying scenes. Analysis indicates challenges in distinguishing subtle or overlapping actions, often due to dataset imbalance. Our results demonstrate the feasibility of leveraging foundational models for real-time bullying detection, while highlighting the need for better balanced datasets and adaptive sampling strategies. Future directions include expanding the dataset, refining annotations, and exploring open-vocabulary recognition techniques.**

***Index Terms*—Bullying detection; action recognition; deep learning; dataset curation; computer vision; machine learning**

I. Introduction

Bullying in schools is a persistent problem in the United States, affecting students' mental health, academic performance, and overall well-being. According to the National Center for Educational Statistics (NCES), approximately 19% of U.S. students aged 12–18 reported being bullied at school during the 2021–22 school year, with incidents most frequently occurring in hallways, classrooms, and cafeterias [2]. While these numbers have declined over the past decade, bullying remains a significant concern, especially in unsupervised areas where intervention is less likely. Notably, the prevalence of reported bullying differs by gender, with 21.8% of female students reporting being bullied compared to 16.7% of male students, as shown in Figure 1.

Traditional school surveillance systems, often consisting of CCTV cameras, offer coverage but lack real-time analytics. These systems typically rely on human monitoring, which is prone to fatigue, limited field of view, and delay in response. As a result, many bullying incidents still go undetected or unaddressed. Recent advancements in deep learning and computer vision, particularly in video-based action recognition—offer the potential to automate detection of such behaviors in real time.

However, current action recognition models are not specifically trained for school bullying detection. Most available datasets focus on general violence or abnormal behavior, lacking the contextual specificity and class diversity required to identify nuanced interactions between students. In addition, there is a shortage of annotated datasets that represent real-world school environments.

In this work, we present a foundation-model based framework for real-time detection of physical bullying behaviors from CCTV footage. We curate a 12-class dataset of over 3,000 video clips from two public datasets: RWF-2000 and UCF-Crime, movies and social media platforms such as YouTube, Facebook, Instagram and X. These clips are annotated with distinct physical actions such as punching, shoving, hair grabbing, and slapping, as well as mixed-label and non-bullying interactions.

We fine tune a vision-language action recognition model using temporally sampled frames (25 per video) and train it with 20 epochs. The resulting model achieves a top 1 accuracy of 43% across all classes and over 90% accuracy on non-bullying scenes. Our analysis reveals high performance on visually distinct actions and lower accuracy on overlapping or subtle behaviors, primarily due to dataset imbalance.

This study demonstrates the feasibility of using AI-based video analytics for bullying detection in schools and highlights key challenges such as motion ambiguity, class granularity, and data imbalance.

II. Related Works

In recent years, video-based violence and bullying detection have become a growing area of interest within computer vision. While traditional CCTV systems lack intelligent analysis capabilities, recent studies have demonstrated the feasibility of machine learning approaches to automate detection of aggressive behaviors.

One of the first large-scale efforts in this domain is the Bullying10K dataset, which uses neuromorphic event-based data to support privacy-preserving bullying recognition [6]. The dataset includes 10,000 segments across multiple tasks such as action classification and temporal localization, using event cameras and synthetic skeleton data. However, the

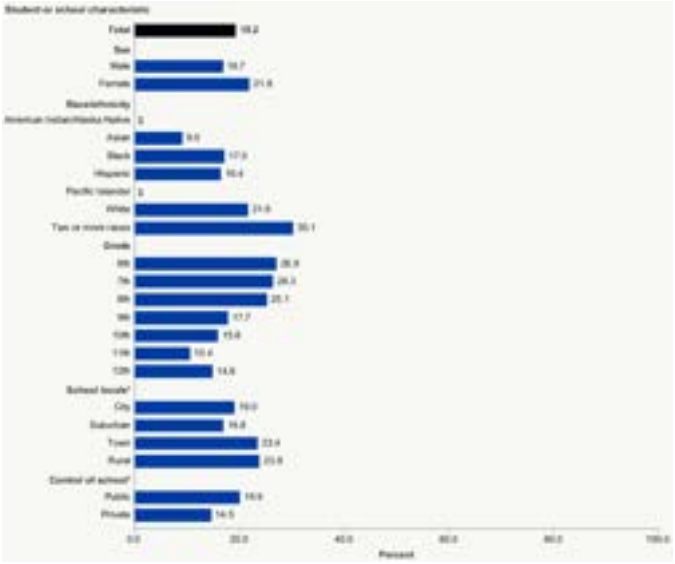


Fig. 1. Percentage of students ages 12–18 who reported being bullied during school, by selected student and school characteristics (2021–22). Source: NCES.

modality is not directly compatible with RGB-based CCTV systems deployed in most real-world school environments.

The authors of [7] proposed a MoveNET-based neural network for detecting physical bullying. Their method first extracts human skeleton keypoints using MoveNET and then feeds these points into an artificial neural network for classification. A dataset of 20 hours of video was curated, containing 13 aggressive behavior classes (e.g., punching, kicking, finger pointing). Their approach achieved up to 98% classification accuracy with relatively low computational cost, making it suitable for real-time deployment. However, the dataset used is not publicly available.

In the area of general video violence detection, several works have employed 3D convolutional neural networks, optical flow representations, or recurrent networks such as LSTM and GRU. Other approaches include multi-stream CNNs [5] and CNN-LSTM hybrids [3], which have shown good performance but often require high computational resources and lack class granularity for school specific bullying detection.

III. Problem Statement

The current challenge is the absence of automated, real-time systems capable of detecting bullying behaviors from school CCTV footage. Most surveillance setups rely on human monitoring, which is slow and often ineffective. Existing action recognition models are not trained on school-specific aggression, and there is limited availability of curated datasets with realistic bullying scenarios.

To address this, we aim to develop a video-based detection framework trained on a custom 12-class bullying dataset. Our goal is to evaluate the performance of fine-tuned foundation models on realistic school interactions and establish a baseline system for future real-time deployment in educational environments.

IV. Data

We constructed a dataset of short video clips depicting physical bullying behaviors in school-like settings by combining content from anomaly detection datasets, violence detection datasets, movies, and social media. The goal was to collect realistic and diverse samples of aggressive interactions, as well as neutral scenarios for contrast. To ensure both action clarity and practical usability, all videos were segmented into 4–8 second clips and manually annotated into predefined action classes, as illustrated in Figure 2.

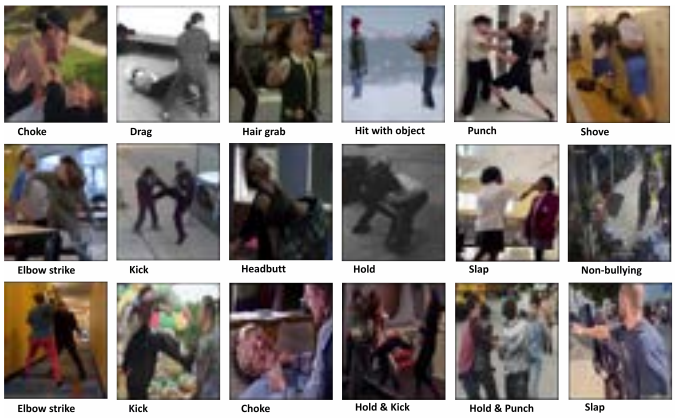


Fig. 2. Example frames from the constructed dataset showing a variety of physical bullying actions.

A. Data Collection

We began by extracting only the anomalous scenes from the UCF-Crime and RWF-2000 datasets, isolating physical interactions such as punching and shoving as distinct classes. All remaining segments were grouped under a single "non-bullying" class to support baseline training and comparison.

Next, we generated a list of school bullying movies and series by prompting ChatGPT. We downloaded subtitles for these videos and used ChatGPT again to identify potential bullying moments. While initially useful, this approach was replaced by a more accurate process: querying YouTube directly for bullying scenes using the recommended movies or series. Some results included compilation videos where multiple physical bullying events were already stitched together, which significantly streamlined the collection process.

Finally, we used social media platforms such as Facebook, X, and Instagram, along with YouTube, to search terms like "school bullying", "hallway fights" and "locker room fights". Relevant clips were screen recorded, segmented into 4-8 second clips using a custom python script, and manually categorized into one of the bullying action classes or a non-bullying class.

B. Dataset

The final dataset consists of 3,081 short video clips, each labeled into one of 11 physical bullying actions, a multi-label category, or a non-bullying class. It included real-world

user recordings and curated anomalies from public datasets. The diversity of sources ensures variation in camera angles, environments, and visual styles, which is essential for training robust action recognition models. As shown in Table I, classes such as *punching* and *shoving* are well-represented, while others such as *headbutt attack* and *elbow strike* are underrepresented. This imbalance reflects the natural frequency of these actions in real-world scenarios and highlights the challenge of learning robust models across all classes.

TABLE I
CLASS DISTRIBUTION AND NUMBER OF VIDEOS

Class ID	Class Type	Number of Videos
0	Choke	90
1	Drag	41
2	Elbow	12
3	Hair grab	48
4	Headbutt	8
5	Hit with object	247
6	Hold	288
7	Kick	296
8	Punch	706
9	Shove	430
10	Slap	119
11	Multi-label	347
12	Non-bullying	449

V. MATERIALS AND METHODS

We used the T2L (Temporal Token Learning) framework [4] for video action classification, as illustrated in Figure 3. T2L is built on top of the CLIP backbone, which provides visual-language pretrained representations. The model incorporates a lightweight temporal token learning mechanism that captures inter-frame dependencies by learning latent temporal structure across sampled frames. This makes T2L particularly effective for short video action recognition tasks, where actions unfold over a few seconds.

The frames from each video clip were resized, normalized, and processed as input tokens to the transformer encoder. We fine-tuned the model for 20 epochs using standard supervised learning with cross-entropy loss across 12 action classes.

The temporal tokens in T2L allow the model to attend to relevant moments while ignoring redundant or irrelevant frames. This is particularly useful in our dataset, where bullying behaviors may occur briefly within neutral interactions. By leveraging the pretrained CLIP backbone and fine-tuning with temporal supervision, T2L captures both appearance and motion cues effectively.

VI. EXPERIMENTAL RESULTS

We evaluated the performance of our model on the curated 12-class dataset after training for 20 epochs using 25 uniformly sampled frames per video. The classification output was compared against ground truth labels to compute accuracy across each class.

The overall top-1 accuracy achieved was approximately 43%, with the model demonstrating strong precision on visually distinct actions and high performance on the non-bullying

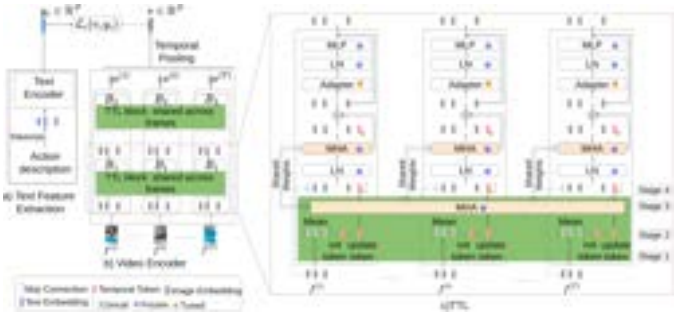


Fig. 3. Overview of the T2L (Temporal Token Learning) architecture used for video action recognition [4].

class, which achieved over 90% accuracy. This suggests that the model is highly effective at distinguishing clearly non-aggressive behavior but faces challenges when differentiating between overlapping or ambiguous actions.

Figure 4 shows the per-class accuracy across all 12 categories. Actions such as *hair grab*, *elbow strike*, and *hit with an object* yielded relatively high classification accuracy, while actions such as *headbutt*, *kick*, and *drag* showed lower performance. These discrepancies are partially attributed to dataset imbalance, with underrepresented classes having limited training examples, and to motion ambiguity, where short physical interactions resemble each other visually.

Analysis of misclassified samples revealed recurring confusion between certain action pairs (see Figure 5). For instance, *slapping* is frequently confused with *punching*, and *shoving* is often misclassified as *aggressive holding*. This highlights the model's sensitivity to subtle motion differences in crowded or occluded settings.

The results emphasize the importance of class balancing and better sampling strategies. While our uniform frame sampling strategy preserved temporal coverage, it may not always capture peak motion segments. Incorporating adaptive frame selection based on motion energy or saliency could improve recognition of short, fast actions. These findings establish a baseline for future models and provide actionable insights for dataset expansion, especially in rare and visually similar classes.



Fig. 4. Per-class accuracy across the 12 categories



Fig. 5. Misclassifications pattern

VII. Conclusion

In this study, we developed a dataset-driven framework for detecting physical bullying behaviors in school environments using CCTV video. A diverse set of short video clips was collected and annotated across 11 physical bullying actions and a non-bullying category. We fine-tuned a vision-language transformer model using the T2L framework on 25-frame sampled segments, achieving a top-1 accuracy of approximately 43%.

Our results show high accuracy in distinguishing clear non-bullying scenes and specific physical actions such as hair grab and hit with an object. However, classification performance declined on subtle or visually overlapping motions revealing challenges related to class imbalance and motion ambiguity. The model frequently confused actions with similar spatial characteristics, such as shove and aggressive holding.

These findings demonstrate the feasibility of using pre-trained transformer-based models for real-time bullying detection in school surveillance footage. Future work will focus on improving model accuracy by expanding the dataset, refining annotations, and incorporating adaptive temporal sampling and open-vocabulary modeling to better handle ambiguous classes.

References

[1] M. Cheng, K. Cai, and M. Li, "RWF-2000: An open large scale video database for violence detection," in *Proc. 25th Int. Conf. on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 4183–4190.

[2] NCES, "*Student Bullying*," The Condition of Education 2024, 2024. [Online]. Available: https://nces.ed.gov/programs/coe/indicator/a10

[3] N. AlDahoul, H. A. Karim, R. Datta, S. Gupta, K. Agrawal, and A. Albunni, "Convolutional Neural Network-Long Short Term Memory Based IoT Node for Violence Detection," in *Proceedings of the 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, IEEE, 2021.

[4] S. Ahmad, S. Chanda, and Y. S. Rawat, "T2L: Efficient Zero-Shot Action Recognition with Temporal Token Learning," *Transactions on Machine Learning Research*, published May 11, 2025. [Online]. Available: https://openreview.net/forum?id=ceve5lB8jV

[5] S. A. Carneiro, G. P. da Silva, S. J. F. Guimaraes, and H. Pedrini, "Fight Detection in Video Sequences Based on Multi-Stream Convolutional Neural Networks," in *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Rio de Janeiro, Brazil, 2019, pp. 8–15. doi: 10.1109/SIBGRAPI.2019.00010.

[6] Y. Dong *et al.*, "Bullying10k: A large-scale neuromorphic dataset towards privacy-preserving bullying recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 1923–1937, 2023.

[7] Z. Kozhamkulova *et al.*, "MoveNET enabled neural network for fast detection of physical bullying in educational institutions," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, 2023.