WVD: A New Synthetic Dataset for Video-based Violence Detection

Muhammad Shahroz Nadeem[1][0000−0001−5835−1602], Virginia N. L. Franqueira[2][0000−0003−1332−9115] Fatih Kurugollu[1][0000−0002−2508−4496], and Xiaojun Zhai[3][0000−0002−1030−8311]

[1] College of Engineering and Technology, University of Derby, Derby, DE22 3AW, United Kingdom {m.nadeem,f.kurugollu}@derby.ac.uk
[2] School of Computing, University of Kent, Canterbury, CT2 7NF, United Kingdom {v.franqueira}@kent.ac.uk
[3] School of Computer Science and Electronics Engineering, University of Essex, Colchester, CO4 3SQ, United Kingdom {xzhai}@essex.ac.uk

Abstract. Violence detection is becoming increasingly relevant in many areas such as for automatic content filtering, video surveillance and law enforcement. Existing datasets and methods discriminate between violent and non-violent scenes based on very abstract definitions of violence. Available datasets, such as "Hockey Fight" and "Movies", only contain fight versus non-fight videos; no weapons are discriminated in them. In this paper, we focus explicitly on weapon-based fighting sequences and propose a new dataset based on the popular action-adventure video game Grand Theft Auto-V (GTA-V). This new dataset is called "Weapon Violence Dataset" (WVD). The choice for a virtual dataset follows a trend which allows creating and labelling as sophisticated and large volume, yet realistic, datasets as possible. Furthermore, WVD also avoids the drawbacks of access to real data and potential implications. To the best of our knowledge no similar dataset, that captures weapon-based violence, exists. The paper evaluates the proposed dataset by utilising local feature descriptors using an SVM classifier. The extracted features are aggregated using the Bag of Visual Word (BoVW) technique to classify weapon-based violence videos. Our results indicate that SURF achieves the best performance.

Keywords: Violence Detection · Dataset · Hot and Cold Weapons · Video Classification · GTA-V · Computer Games · WVD

1 Introduction

One of the fundamental challenges for building violence detection systems is the subjective nature of violence [2]. Violence can be expressed in verbal, physical and physiological forms. Particularly, it is a common observation that in acts of deliberate physical violence, weapons play an important role in causing harm to others. These weapons are broadly classified as hot (containing gunpowder, which cause fire and explosion) or cold (do not contain gunpowder) weapons.

Current work in violence detection, only focuses on discriminating violent scenes against nonviolent. MediaEval has provided a large scale authoritative benchmark for violence detection. However, it is based on Hollywood movies, thus, categorised as containing staged violent videos. Staged sequences lacks the human behaviour and components exhibited during fights in real world settings. Moreover, it is designed to discriminate videos containing violence against non violence. Nievas et al. [5] proposed the Movies and Hockey datasets. These datasets contain fight sequences taken from Hollywood movies and the National Hockey League but do not contain any weapons usage. Further, YouTube is a prominent source to gather real-world video data. However, their policies restrict uploading content which involves violence, gore or disturbing content which can incite others to commit violent acts [4]. These factors greatly restricts gathering data that captures weapon-related violence. Therefore, these existing datasets 1) lack the presence of any type of weapon in fights, 2) contain videos taken from Hollywood movies which are staged, 3) have scalability and ethical issues.

In this paper, these gaps are addressed through a synthetically generated dataset of violent sequences for hot and cold weapons using the photo-realistic game Grand Theft Auto-V (GTA-V), named as "Weapon Violence Dataset" (WVD). The dataset is available on request from the authors. Recently, in computer vision simulated environments, games and frameworks are been used for designing, labelling and gathering data. This virtual data is then used to train algorithms. Autonomous vehicle research is a prime example where synthetic data is utilised [3] [7] [1]. GTA-V has also been used for synthetic data generation [6] [4]. The contribution of the paper include a novel video-based dataset for weapon-based violence. Our dataset contain weapon based violent interactions, governed by game's AI thus are not staged. Further, it solves the ethical or moral implication problem. We evaluated the proposed dataset with an SVM classifier using local feature descriptors which include: Histogram Of Gradients (HOG), Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF) and Oriented FAST and Rotated BRIEF (ORB). The low-level features extracted from these are passed to Bag of Visual Words (BoVW) technique used in image recognition to aggregate low-level feature to develop high-level features, on the WVD dataset. Our results indicate that features extracted by SURF are superior in-comparison to the experimented feature extractors. The remaining of the paper is organised as follows.

Section 2 explains the characteristics of our WVD dataset using BoVW. In Section 3 we show experimental results on the WVD dataset. Finally, Section 4 concludes the paper where we highlight our main contributions and future research directions and goals.

2 Hot vs Cold Weapons Violence Dataset

GTA-V is an open world game, which gives us the freedom to design violent scenarios for violence detection. Several characteristics of this game made it an ideal candidate for generation of the dataset. Firstly, GTA-V has a strong "mod"

[4] https://support.google.com/youtube/answer/2802008?hl=en-GB

Table 1: The list of weapons used in the dataset.

Hot Weapons		
AP Pistol (OTs-33 Pernach or HK Mark 23)	Combat Pistol (HK P2000)	Pump Shotgun (M590A1)
Bullpup Rifle (QBZ-95)	Carbine Rifle (HK416, LR-300)	Assault Shotgun (UTAS UTS-15)
Micro SMG (IMI Uzi)	SMG (HK MP5)	MG (PK)
Combat MG (M249)		
Cold Weapons		
Baseball Bat	Broken Bottle	Crowbar
Hammer	Hatchet	Knife
Pipe Wrench	Machete	Golf Club

(modification) support and community which generate scripts, with them it is possible to change different aspects of the game. Secondly, GTA-V allows to capture the design scenarios under different times of the day, multiple camera angles and a huge array of different weapons, vehicles and objects. Moreover, the ability to place and edit the appearance, fighting styles, stances and health of the Non-Player Characters (NPC). Thirdly, NPCs can fight independently without human supervision or involvement. These factors make GTA-V an ideal candidate to be chosen as a virtual platform for weapon based violence data generation.

The proposed WVD was generated in a three-step process, the first step includes designing violent scenarios through mods which use 'ScriptHookV' and 'ScriptHookDotNet'. Each scenario consisted of two NPCs, which were assigned the roles of the 'aggressor' and the 'victim'. In every scenario, the aggressor NPC is equipped with a weapon (either hot or cold), their combat style was set as either aggressive or defensive. However, in case of hot weapon fights the combat style was set to stationary. The reason for this was to avoid the NPC moving away from the field of capture. In GTA-V's combat engine, we observed that NPCs maintained a specific safe distance while fighting with hot weapons in order to protect themselves. They also ran towards the nearest object in order to take cover. These safety features made capturing the hot violent sequences difficult. As our goal was to focus on fights where the intention of the aggressor is to kill or inflict fatal wounds, also the victim is unarmed this removes the necessity of maintaining a safe distance or taking cover. Due to this, the fighting style stationary was fixed. Depicting the roles of aggressor and victim. The victim NPC were unarmed and were assigned different combat styles however, they had to rely just on their fists to fight the aggressor. The scenarios were set up in urban, industrial and rural settings. The weapons utilised are mentioned in Table 1.

Once the scenarios were set up, the second step was to run these scenarios. It must be noted that no human supervision or assistance was used during the duration of the fights. This meant that we had no control over the fight and its end result. The mechanics of the game set the rules and the outcome of fights.

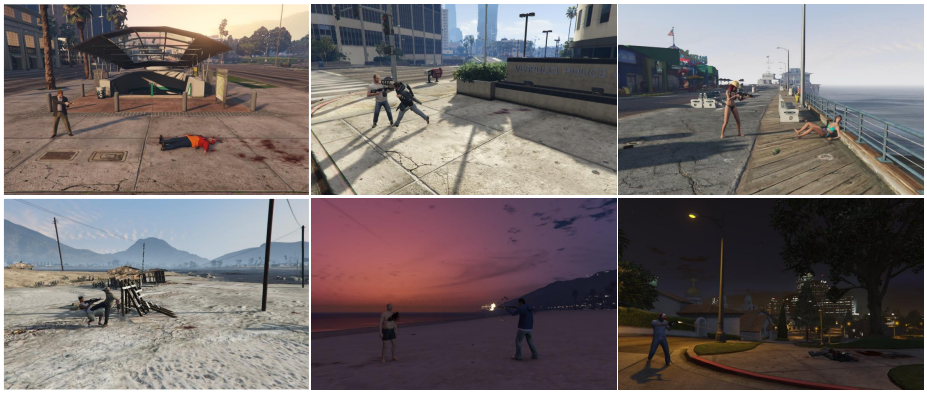

Fig. 1: Frame sequences capturing fights with hot weapons. The captured images exhibit the presence of blood, gun fire and occasionally close quarter combat.

This meant that running a particular scenario multiple times would result in different actions taken by both the aggressor and victim NPC. Moreover, we had no control over the positioning of the NPCs during the fights. Due to this, NPCs' had variations in the position and depth of their fights. However, due to the visible advantages of the weapon, that the aggressor NPC had of extended range and damage, resulted in aggressor winning the fight. These scenarios were set up under different times of the day which include Morning, Midday, Midnight, Dusk, Afternoon and Sunset. The final step required human supervision where the captured sequence was viewed by human observers for anomalies. For each captured video, unwanted frames were removed. Figure 1 shows a sample of frames captured from hot weapon fights. Here we can observe the visible presence of blood and gunfire. While Figure 2 illustrates cold weapons with 'swinging' and 'shoving' actions being performed against the victim.

3 Experimental Evaluation

In this paper, we selected the appearance-based visual features following the example of Nievas et al. [5] on the WVD dataset. For low-level feature extraction, the descriptors selected include HOG, SIFT, SURF and ORB. The features extracted from these descriptors are used to generate the vocabulary using K-Means. Finally, an SVM classifier is trained to distinguish between Hot and Cold fights. For experimental purposes, the frames were resized from a $800 \times 500 \times 3$ dimensional frame to $400 \times 250 \times 3$. This resulted in two variants of our dataset with 5.34 GB and 1.52 GB in size. We utilised the reduced set for training. A total of 40,845 frames were extracted from the videos in WVD. This resulted in multiple frames from each scenario with different levels of NPC depth, weapon occlusion and fight styles.

The WVD dataset was divided into 70% training and 30% testing sets. Each of the methods took approximately two to three days to train on an Intel Xeon W-2123 3.60 GHz with 64 GB of RAM on the resized version of the dataset.

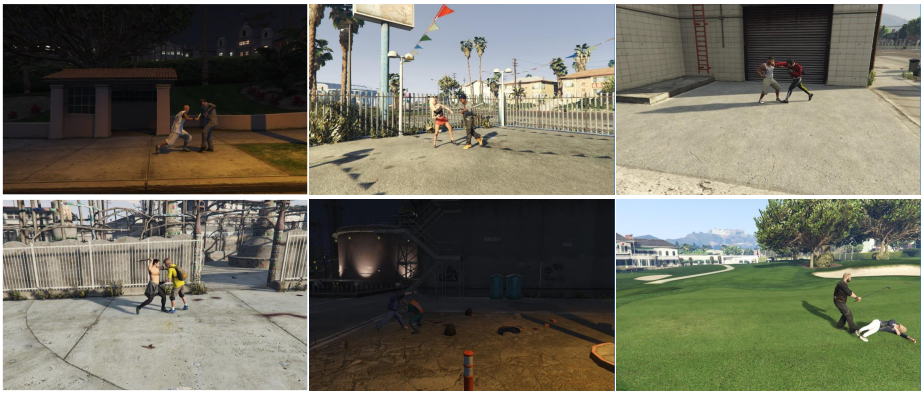

Fig. 2: Frame sequences capturing fights with cold weapons. The captured images exhibit the presence of swinging, shoving and occasionally dodging the attacks from cold weapons

First, low-level features are extracted from individual frames by local feature descriptors. These low-level features are aggregated to develop higher-level feature representation through clustering techniques. Afterward, the SVM classifier is trained upon these high-level features. As HOG, SIFT, SURF and ORB all extract low-level features. Thus combining them together would result in high number of features, which would have increased the computational complexity. Further, all such features would be aggregated during BoVW representation. The goal was to evaluate the quality of the features rather than the quantity. Due to this reason, we did not combined these feature extractors.

Based on experimental evaluation, SURF performed better than all the other feature extractors. SIFT and HOG had similar performance, while with slight variations in F1 scores. However, SIFT has lower precision and high recall for hot weapon fights, compared to HOG, which has higher precision and lower recall. SIFT was much more computationally friendlier in comparison to HOG, which was found to be the most memory hungry. ORB performed poorly amongst all the local feature descriptors. Results indicate that SURF has the highest precision while HOG has the highest recall for cold fights. Overall, hot weapon fights have a lower precision rate in comparison to cold weapon fights as shown in Table 2. This shows that the classifiers are more confused when it comes to identifying hot weapon fights.

4 Conclusion and Future Work

In this paper, we presented a new synthetic virtual dataset called WVD built for hot and cold weapon-based violence using the photo-realistic game GTA-V. The dataset focuses explicitly on fights with different weapon types between individuals, something not discriminated in authoritative datasets for violence detection such as the “Hockey Fight” and “Movies”. To the best of our knowledge, this is the first synthetic dataset for this problem which is not based on Hollywood

Table 2: The reported precision, recall, F1-scores and accuracy (ACC) are shown for the selected feature extractors.

SIFT	Precision	Recall	F1 Score
Cold	0.90	0.79	0.84
Hot	0.68	0.84	0.75
ACC			0.81

SURF	Precision	Recall	F1 Score
Cold	0.92	0.91	0.91
Hot	0.84	0.85	0.84
ACC			0.89

ORB	Precision	Recall	F1 Score
Cold	0.87	0.76	0.81
Hot	0.65	0.80	0.72
ACC			0.78

HOG	Precision	Recall	F1 Score
Cold	0.82	0.92	0.86
Hot	0.78	0.65	0.71
ACC			0.81

movies and contains a range of weapon-based violence. Our approach is scalable and does not suffer from any ethical implications associated with violence. Further experiments are performed using the BoVW approaches using four different local feature descriptors. Our experimental results indicate that SURF is by far the better feature descriptor achieving a F1 score of 0.91 for cold and 0.84 for hot weapon fights. This means that WVD retains enough visual information. In this paper, we only focused on appearance-based visual features. To further evaluate the dataset, we aim to combine motion-based features with visual features. Furthermore, we also aim to diversify our dataset by adding 'no-violence' class as a control group and capturing scenarios for these classes under different camera angles and multiple weather settings.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019)
2. Demarty, C.H., Penet, C., Soleymani, M., Gravier, G.: Vsd, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation. Multimedia Tools and Applications **74**(17), 7379–7404 (2015)
3. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. arXiv preprint arXiv:1711.03938 (2017)
4. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2821–2830 (2018)
5. Nievas, E.B., Suarez, O.D., García, G.B., Sukthankar, R.: Violence detection in video using computer vision techniques. In: International conference on Computer analysis of images and patterns. pp. 332–339. Springer (2011)
6. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European Conference on Computer Vision. pp. 102–118. Springer (2016)
7. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016)