Brutality detection and rendering of brutal frames

R Madana Mohana[1], Paramjeet Singh[1*], Vishal Kumar[1], and Sohail Shariff[1]

Department of Artificial Intelligence and Data Science, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India

Abstract. The popularity of anime is increasing exponentially in every part of the world due to its unique storyline, nonstop entertainment, fights, and similar type of content that can hold viewers and keeps them at the edge of their seats. However, with the increase of popularity in anime there has also been an exponential increase in violence and brutality in anime videos. Violent scenes have become much more common in anime videos when compared to generic cinema. This survey paper presents a comprehensive view on the detection of violence in movies and different scenarios using various techniques. Most commonly to automate detection of violence, machine learning is used for training the machine to detect violence. Convolution neural networks (CNN) are used very commonly to understand image pattern recognition with high accuracy. Moreover, use of other different methods such as LSTM and Markov models are also used to detect violence. The main goals kept in mind while working is to detect violence with high accuracy and to use less computation or to perform the action at a high-speed rate.

1 Introduction

In the good old days of cinema, the early days brought the magic of motion pictures to audiences around the world. From silent films to the introduction of sound, from black & white to colour, cinema has evolved tremendously over the years. The impact of cinema on people has been significant. Movies have the power to transport viewers to different worlds and inspire creativity and imagination. The influence of cinema on popularity and society is evident in the way films can shape trends, influence fashion, and impact social and political attitudes.

In recent years, anime has gained immense popularity, particularly among younger generations. The popularity of anime can be attributed to its unique style, compelling storylines, and characters that resonate with audiences.

The violence portrayed in anime is more intense and graphic than what is seen in mainstream movies or TV shows, and that it can have a significant impact on viewers, particularly younger audiences.[1] Studies have shown that exposure to violence in media, including anime, can be sensitive to viewers and lead to aggressive behavior. Additionally, the immersive nature of anime can make violent scenes feel more real and intense, which may have a stronger impact on viewers.

* Corresponding author: paramjeets0601@gmail.com

It is important to note that not all anime contains graphic violence, and many series explore complex themes and ideas without relying on violent imagery. It is up to individual viewers to make informed decisions about the media they consume and to understand the potential impact it may have on their attitudes and behaviour. Additionally, parents and guardians should monitor their children's media consumption and have open discussions about the content they are exposed to.

It's clear how and why [2] anime has gained so much popularity over the years and one of the major reasons is delivering high quality content in less time is one of the major reasons, though violence content in anime is a common theme that often shows violence and aggression in various forms and due to recent effects of the COVID-19, many got time to spare and relax at here homes due to this the digital media had skyrocketed sky d gained more popularity with streaming platforms like Amazon prime, Netflix, Hotstar, Aha, and many non-live actions movies and series got popular. While these Anime visuals are often used to create a dramatic effect, they can also be unsettling and disturbing to viewers. To address these concerns, researchers have begun exploring methods for detecting and classifying brutality in anime videos.

The development of such methods is crucial in helping to identify potentially harmful content and providing a safer viewing experience for viewers. These methods typically involve using machine learning algorithms to analyse the visual and auditory features of anime videos and classify them into categories. One major challenge in developing these methods [3] is the subjective nature of brutality, as it can be difficult to define and quantify. Another challenge is the need to balance accuracy with efficiency, as these methods must be able to quickly process enormous amounts of data in real-time. Despite these challenges, the development of brutality detection methods is a crucial step toward ensuring the safety and well-being of anime viewers.

Convolutional Neural Networks (CNNs) [4] are a foundational tool for image classification. They begin by preparing and preprocessing image datasets, followed by selecting an appropriate architecture, such as LeNet [5], VGG [6], or ResNet [7]. Initialising the model, often with pre-trained weights, saves time and computational resources. The CNN is then trained on a labelled dataset, learning to recognize patterns and features that distinguish different image classes through backpropagation and optimization algorithms. Regular evaluation on a validation set ensures the model's accuracy and helps prevent overfitting. Fine-tuning may be applied, and the performance is rigorously tested on a test set to estimate real-world capabilities. Once satisfactory, the model can be deployed for image classification in various applications. Continuous monitoring and potential retraining with new data are essential to maintain the model's accuracy over time. CNNs' ability to automatically learn intricate image features makes them stand out in the world of computer vision.

The use of machine learning for brutality detection in anime videos involves training models using annotated datasets of anime videos. These datasets are often annotated by human experts who watch the videos and label them according to various categories, such as violence, bloodshed, and gore.

2 Literature Review

In Paper [8], Markovian approach is used to predict violence scenes from movies. Here 2 Markovian models are used, first one is used for emotions from text and second works is

used for emotions from frames. It suggests two options of either cut or blur of violent frames. Cascade Classifier [9] is used to model for face expression. The dataset is of 2000 clips of total size of 2GB is used with input dimension of 1920x10080. Markov chain model is used to get the probability between emotions with the increase of number of emotions. The accuracy increases with increase in number of emotions. Last 400 emotions from the prevalent scene give accuracy of 0.8 where the last 800 emotions give an accuracy of 0.93.

In Paper [10], Focus on elimination of unnecessary redundant frames are done which improves accuracy. For this ConvLSTM[11] based violence detection system (ECLVDS) is used. Here the hockey fight data set is used where the vgg19 model is used for feature extraction. This is then given to ConvLSTM which consists of 128 filters. Algorithms such as Hu Moments are used for shape features. Furthermore, k means [12] is used to find the final cluster of features. The accuracy achieved in this paper is 98.90. Using LSTM instead ConvLSTM gives accuracy reduced by 1.8 %. This shows that ConvLSTM can be better when compared to LSTM, but this might also increase the complexity in the model.

In Paper [13], The study uses Hockey Fights and Movies Dataset. This contains 500 violent and nonviolent each. Each video has a duration of about 2 seconds. There are 3 representations, first is spatial temporal that relies on bidirectional convolutional LSTM second is densely connected network based on 3D convolution and 3rd is multi modal detection algorithm for weak supervision though these 3 combined obtained exceptionally outstanding results but can be improved a lot, in every aspect of performance.

In Paper [14], An effort to solve the two common problems in the CNN model is done. That is a decrease in accuracy and overfitting of models. Here Resnet along with batch normalisation is used. ResNet solves this problem by retaining network performance even accuracy reaches a saturation point. Here the data is given into 5 distinct categories. The difference between accuracy of ResNet50 and ResNet101 is only 0.1%. This shows that ResNet can be particularly useful if there is plenty of data available. But a point, point increase in the number of layers of ResNet architecture does not give comparable results.

In Paper [15], Datasets: Hockey Fights dataset is 121.271 seconds using SVM with a polynomial kernel. The polynomial kernel testing is the fastest. It takes 0.6 seconds to classify a video from hockey datasets, Transform (DWT) for feature extraction and SVM[16] for classification And PCA[17] for feature extraction, before carrying out the PCA process, the data is first normalized .To classify violent or nonviolent frames we used dot for feature extraction and SVM for classification : SVM is best when it comes to one dimensional data but image and frames its performance drops drastically.

In Paper [18], An effort to perform real time violence detectors is done where work is done on both speed and accuracy. Here CNN is used to extract Feature and LSTM as a temporal relational learning method. This approach combines usage of VGG19 followed by LSTM. Output of VGG19 is given to 40 cell LSTM. This paper uses a combination of Hockey movies and violent crowds. a total of 896 videos are used. Accuracy obtained is 98% and the frame speed is 131 per second. This paper highlights the importance of fusion of CNN model along with LSTM which can be used to detect violence in combination of data sets. However, the amount of data taken here is small in quantity. VGG as feature Extractor and LSTM to work with series can be used to improve the accuracy.

This paper [19] introduces VNet, a specialised Deep Violent Flow Network designed for violence detection in video sequences, with a focus on abnormal velocity patterns. The model

demonstrates excellent performance on distinct datasets: achieving 99% accuracy for movies, 94% for crowd scenarios, and 98% for hockey videos. The dataset comprises 1000 hockey videos, 500 of which involve violent actions, each with 50 frames at 288x360 p resolution. Additionally, the dataset includes 246 crowd videos with 50 frames at 240x320 pixel resolution, and 200 movies, each with 50 frames at 250x360 pixel resolution. The ViNet architecture incorporates Vgg16 and Inception V3, both known for efficacy in image analysis, along with four hidden layers. This combination proves effective in detecting forged images. Overall, the study highlights ViNet as a powerful tool for violence detection in video sequences, highlighting the importance of considering abnormal velocity patterns as a key feature for accurate classification.

The Paper [20], presents a robust violence detection system leveraging ResNet50 in tandem with a single shot detector (SSD) to analyse video streams in real-time. The integration of ResNet50, a powerful convolutional neural network, with SSD enhances the system's ability to identify intense incidents promptly. The dataset chosen for evaluation, the Hockey Fights Dataset, provides a diverse range of scenarios, contributing to the model's adaptability. The achieved results of an averaged precision of 0.83 and accuracy of 0.846 demonstrate the system's high efficacy in accurately detecting violent incidents. The proposed system holds significant potential for enhancing public safety by enabling swift responses to critical situations.

The Paper [21], introduces a novel violence detection approach leveraging three-dimensional Convolutional Neural Networks (3D CNNs) to extract features across consecutive frames effectively. The study incorporates prominent pre-trained deep learning architectures—VGG16, VGG19, and ResNet50—from the ImageNet dataset, enhancing the models' capability to discern meaningful patterns. Evaluation encompasses diverse datasets, including Hockey (1000 images at 360x288 resolution) and Violent Flow (246 images with variable resolutions), demonstrating the versatility and effectiveness of the proposed methodology in extracting pertinent features from sequential video data. This work not only advances violence detection techniques but also underscores the adaptability and efficacy of 3D CNNs in analysing sequential video data.

This research [22] explores violence detection using pre-trained deep learning models, ResNet50 and VGG16, applied to a diverse range of datasets including Hockey, Movie Fight, Violent Flow, and Real-life Violent Detection. All frames are standardised to 224×224×3 resolution. The proposed methods demonstrate high accuracy across various scenarios, with ResNet50+NN achieving accuracies of 96%, 94%, 100%, and 97%, while VGG16+NN achieves 95.50%, 96%, 100%, and 96% accuracy for Hockey, Violent Flow, Movie Fight, and Real-life Violent Detection, respectively. Particularly notable is the Violent Flow Dataset's impressive 98% accuracy. This research highlights the efficacy of pre-trained models in violence detection, offering valuable insights for real-world applications in security and public safety.

The Paper [23] model is trained on a Mix dataset. A dataset formed by combining different datasets. The model is compared with state-of-the-art models. The test dataset is 20 videos from YouTube. The main model performs better than the state-of-the-art models even if there is 60 frames less for a 1 second video clip.

The model was able to perform better with the introduction of keyframing with respect to both accuracy and lesser computational power. The model is trained to reduce overfitting. The performance is better with less frames. The proposed model with keyframing performs

19% better than the generic interval sampled. And the keyframe extracted ResNet50 – ConvLSTM does 19% better than CNN BiLSTM. The key framing approach as a preprocessing method can be used and tested on other violence detection methods.

3 Proposed Methodology

The research process aims to build a model which can be used to detect violence in the anime videos. We tend to use CNN models to detect if a frame contains violence or not. This will be done for all the frames of a video under consideration as an input.

The working project can be divided into 3 parts. First is conversion of video which is taken as input into number of frames. Even though number of frames depend on the video and a fixed number cannot be taken, most anime videos in today's scenario are made from 24 frames. So, the division of video into frames is proposed which will be stored in a sequential order. After achieving list of frames, the data will be given as input to the model. Then the list of all frames which are identified as brutal is collected and stored in an order. It is important to change those brutal frames into blur frames so that the violent part in the Anime can get hidden. This brings the frames which needed to be re rendered in an order. Using different libraries, the frames can be rendered into the video with same duration.

To create the CNN model first the brutal and non-brutal images at classified manually. For this large quantity of videos are taken and. The videos are divided into multiple number of frames. The frames are divided manually for the training data set and image pattern recognition model is built on the training data set. Accuracy and other metrics for the success of the model is determined and noted.

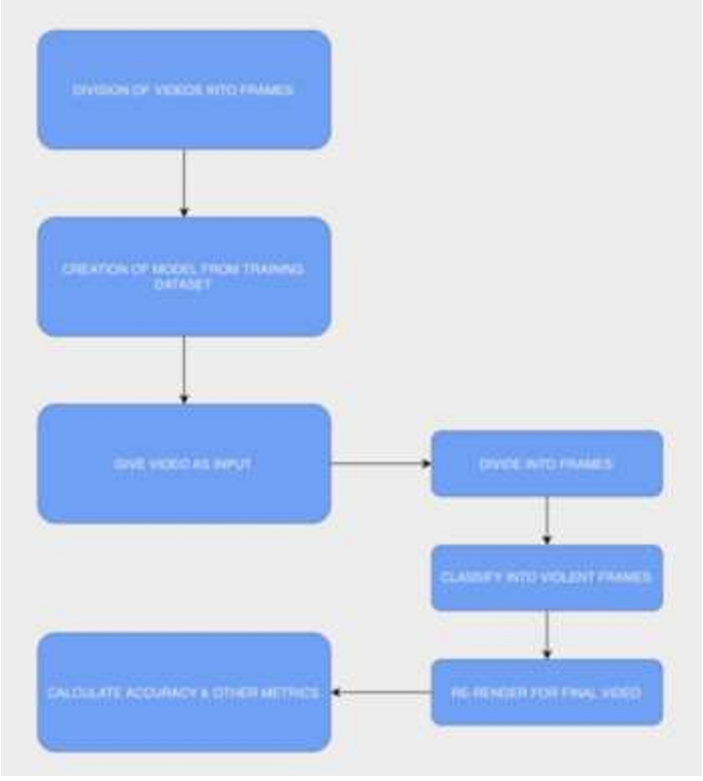

Fig. 1. Structure of the model

References

1. Worth KA, Gibson Chambers J, Nassau DH, Rakhra BK, Sargent JD. Exposure of US adolescents to extremely violent movies. Pediatrics. 2008 Aug 1;122(2):306-12.
2. Napier SJ. Why Anime?. InAnime from Akira to Princess Mononoke: Experiencing Contemporary Japanese Animation 2001 (pp. 1-13). New York: Palgrave Macmillan US.
3. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of big Data. 2021 Dec;8:1-74.
4. 4. Wu J. Introduction to convolutional neural networks. National Key Lab for Novel Software Technology. Nanjing University. China. 2017 May 1;5(23):495.
5. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998 Nov;86(11):2278-324.
6. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).
8. Saad K, El-Ghandour M, Raafat A, Ahmed R, Amer E. A Markov model-based approach for predicting violence scenes from movies. In2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) 2022 May 8 (pp. 262-267). IEEE.
9. Alpaydin E, Kaynak C. Cascading classifiers. Kybernetika. 1998;34(4):369-74.
10. Parui SK, Biswas SK, Das S, Chakraborty M, Purkayastha B. An Efficient Violence Detection System from Video Clips using ConvLSTM and Keyframe Extraction. In2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON) 2023 Feb 10 (pp. 1-5). IEEE.
11. Understanding LSTM- a tutorial into Long Short-Term Memory Recurrent Neural Networks
12. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics). 1979 Jan 1;28(1):100-8.
13. Su M, Zhang C, Tong Y, Liang B, Ma S, Wang J. Deep learning in video violence detection. In2021 international conference on computer technology and media convergence design (CTMCD) 2021 Apr 23 (pp. 267-272). IEEE.
14. Li H, Chen F, Mou Y, Li Y. Mixed Real-Life and Anime Harmful Images Classification Using Deep Residual Neural Networks and Migration Learning. In2021 6th International Conference on Image, Vision and Computing (ICIVC) 2021 Jul 23 (pp. 160-164). IEEE.
15. Su M, Zhang C, Tong Y, Liang B, Ma S, Wang J. Deep learning in video violence detection. In2021 international conference on computer technology and media convergence design (CTMCD) 2021 Apr 23 (pp. 267-272). IEEE.
16. Noble WS. What is a support vector machine?. Nature biotechnology. 2006 Dec 1;24(12):1565-7.
17. Yang J, Zhang D, Frangi AF, Yang JY. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE transactions on pattern analysis and machine intelligence. 2004 Jan;26(1):131-7.

18. Abdali AM, Al-Tuma RF. Robust real-time violence detection in video using cnn and lstm. In2019 2nd Scientific Conference of Computer Sciences (SCCS) 2019 Mar 27 (pp. 104-108). IEEE.
19. Ehsan TZ, Mohtavipour SM. VioNet: a deep violent flow network for violence detection in video sequences. In2020 11th International Conference on Information and Knowledge Technology (IKT) 2020 Dec 22 (pp. 82-88). IEEE.
20. Shripriya C, Akshaya J, Sowmya R, Poonkodi M. Violence Detection System Using Resnet. In2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) 2021 Dec 2 (pp. 1067-1072). IEEE.
21. Honarjoo N, Abdari A, Mansouri A. Violence detection using three-dimensional convolutional networks. In2021 12th International Conference on Information and Knowledge Technology (IKT) 2021 Dec 14 (pp. 188-191). IEEE.
22. Honarjoo N, Abdari A, Mansouri A. Violence detection using pre-trained models. In2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA) 2021 Apr 28 (pp. 1-4). IEEE.
23. Wintarti A, Puspitasari RD, Imah EM. Violent Videos Classification Using Wavelet and Support Vector Machine. In2022 International Conference on ICT for Smart Society (ICISS) 2022 Aug 10 (pp. 001-005). IEEE.