Real-Time Violence and Hate-speech Detection in Video for Kid

Dung Hoang Dao[1,2], Ngan Thi-Kim Huynh[1,2], Tinh Thanh Ho[1,3], and Hop Trong Do[1,2]

[1] Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City,Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] Faculty of Computer Science, University of Information Technology, Ho Chi Minh City,Vietnam
{21521972, 21520357, 20520813}@gm.uit.edu.vn, hopdt@uit.edu.vn

Abstract. Currently, the age at which children access the Internet and social media is decreasing, leading to concerns about their cognitive development and overall growth. Exposing children who lack sufficient awareness to videos containing inappropriate content, such as violence or hate speech, can negatively impact their development and educational progress. In this study, we propose a system for Real-time Violence and Hate-speech Detection in Video for Children using three models: Assembly AI, MobileNetV2, and a Hybrid CNN-LSTM-Conv3D for the tasks of hate speech detection, blood detection, and violent action detection, respectively. We focus on improving the accuracy of inappropriate content detection and optimizing execution time. Experiments conducted on 50 short videos collected from YouTube and manually labeled for child-appropriateness yielded 90% accuracy and 85% recall, demonstrating the system's effective prediction of videos unsuitable for children. We also integrated Apache Spark and Kafka to support real-time video processing, achieving an average of 40 frames per second.

Keywords: Real-time · Violence Detection· Multi-classification · Apache Spark · Kafka

1 Introduction

Vietnam currently has over 76 million Facebook users. TikTok, which entered Vietnam just three years ago, already has 50 million users, ranking 6th among the top 10 countries with the highest TikTok user base globally. According to Statista, the average age of TikTok users ranges from 12 to 24 years old [3], with 71% of users reporting that the platform significantly impacts their lives. A survey by the Ministry of Labor, War Invalids and Social Affairs shows that Vietnamese children use social media for 5-7 hours daily; however, only 36% of children are educated about online safety [4].

In Vietnam, many children have lost their lives by following bizarre challenges. Other harmful content may not take lives directly but is "killing" the

souls and thoughts of many young children daily. This includes cyberbullying. Many former offenders now live stream challenging others to violent confrontations. Concerningly, these clips receive hundreds of thousands of views and likes. Many children even idolize online violence. YouTube, one of the most popular social media platforms, has developed a section called YouTube Kids for child-appropriate content. Unfortunately, harmful content is still mistakenly displayed on this platform [1], [2]. This is not to mention platforms that do not curate content before dissemination, which poses a challenge for controlling harmful content for children today.

Currently, a large number of videos are uploaded without content moderation on open social media platforms like YouTube, TikTok, etc., posing potential risks for children to access harmful and violent content. Targeting children who spend more time watching videos than other age groups, along with social media algorithms, some channels have created inappropriate yet attractive content to exploit children in spreading misinformation. This hinders children's development and educational progress. Therefore, we see the necessity of creating a system to detect inappropriate content such as violence, blood, or hate speech.

By leveraging the power of Deep Learning and taking advantage of Apache Kafka and Apache Spark, we develop a system classifier to automatically detect violence and hate speech in videos. In this research, we focus on detecting videos with violent content such as fighting and brawling. Additionally, we simultaneously detect harmful bloody content. These not only appear alongside violent content but are also mixed with non-harmful content disguised as child-friendly. Furthermore, audio in videos is also a factor affecting children's psychology and behavior. It is evident that children at a young age learn quickly, and early exposure to profane language and hate speech can lead to severe consequences. Our contributions to this research are as follows:

- **Proposing a new model with high performance**: Through experimentation, we compare our custom combined model with previously applied popular methods. This demonstrates that our model's performance is superior to most other methods.
- **Combining multi-class detection and classification in videos**: The system allows for the detection of bloody features in videos, classification of violent and non-violent content, and simultaneously distinguishes whether the audio in the video contains hate speech, thanks to three parallel deep learning models.
- **Real-time model deployment**: By utilizing Apache Kafka and Apache Spark, the three models are executed in parallel in real-time. This enables timely detection and prevention of harmful content on social media.

2 Related Work

2.1 Violence dectection in Video for Kid

Datta et al, [10] tried to detect person-on-person violence in videos which involve only fist fighting, kicking, hitting with objects etc., by analyzing violence at

object level rather than at the scene level as most approaches do. Deniz et al, [11] introduced a method for detecting violence in videos using extreme acceleration patterns as the key feature. This method is 15 times faster than state-of-the-art action recognition systems and highly accurate in identifying scenes with fights. It is ideal for real-time violence detection, balancing both speed and accuracy. The approach involves comparing the power spectrum of consecutive frames to detect sudden motion, classifying scenes as violent or non-violent based on the motion detected.

The initial attempt to detect violence using both audio and visual cues is by Nam et al, [8]. In their work, both the audio and visual features are exploited to detect violent scenes and generate indexes so as to allow for content-based searching of videos. In the work by Lin and Wang, [9] a video sequence is divided into shots and for each shot both the audio and video features in it are classified to be violent or non-violent and the outputs are combined using co-training. Tirupattur et cal, [12] had built a system which automatically detects violence not only in Hollywood movies, but also in videos from the video-sharing websites like YouTube and Facebook. In this work, an attempt is made to also detect the category of violence in a video, which was not addressed by earlier approaches. The categories of violence which are targeted in this work are the presence of blood, presence of cold arms, explosions, fights, screams, presence of fire, firearms, and gunshots.

Jahlan et al. [17] proposed a novel method to detect violence using a fusion technique. The performance of this method showed significant improvement when compared to other approaches. Instead of using a single feature extraction layer, this study employed two pre-trained feature extraction layers and utilized a fusion technique to combine them. Subsequently, the results were classified based on the probability of this combination. This demonstrates the importance of using appropriate feature extraction methods. Similarly, Wen-Feng Pang et al. [18] proposed a neural network containing three modules for fusing multimodal information (visual and audio features). This architecture became the state-of-the-art (SOTA) for the XD-Violence dataset.

2.2 Real-Time Violence Detection in Video

Vicente et al. [5] proposed a Violent Flow (ViF) variation using Horn-Schunck instead of Iterative Reweighted Least Squares (IRLS) as the optical flow algorithm and Support Vector Machine (SVM) as a classifier. Evaluation on the Hockey dataset showed better performance; however, it performed worse when assessed on Movies and Crowded datasets. Tao Zhang et al. [6] proposed a Gaussian Model of Optical Flows that, when passed to the linear classifier, identifies regions where violence is inferred. Almamon et al. [7] proposed a pseudo real-time violence detection system, which processes a video, with or without audio, using a CNN + LSTM approach, and alerts when violent activities are detected. This reveals that the real-time systems published for this task have not yet applied the robustness and flexibility of Kafka and Spark.

3 Dataset

To serve the experiment, we have selected three high-resolution datasets for violence recognition tasks, with varying quality, pixel count, and length across the datasets. Hockey [14] consists of fighting scenes in games; Movies [14] includes fighting and violent scenes cut from films; and RLVs [13] is collected from YouTube, containing violence videos with many real street fight situations in various environments and conditions. Additionally, non-violence videos are collected from many different human actions like sports, eating, walking, etc. The specifications of these datasets are described in 1.

To objectively evaluate the system in real-world situations, we have collected 50 short video clips from YouTube and manually labeled them. All of these are in the test set, of which 25 videos are labeled as unsafe for children, including a combination of cases with labels for violent actions, bloody, and hate speech audio.

Dataset	Total videos	Labelled Violent	Labelled Non-Violent	Average Frames
Hockey fights [14]	1000	500	500	50
Movies Fight [14]	200	100	100	50
RLVs [13]	2000	1000	1000	100
Youtube clip (Ours)	50	25	25	50

Table 1: Datasets features

3.1 Data Preprocessing

In preparing the input data, videos are first extracted into sequences of frames. Each frame is taken from the video to form a sequence. The frames are divided into fixed batches. This facilitates easier processing and aligns with the available computational power. The input for the encoding model consists of images generated by subtracting pixels between adjacent frames. This aims to incorporate spatial motion in the video rather than using raw pixels from each frame. Finally, the images are resized to 224 pixels by extracting a square region of 224x224 pixels from the input image.

4 Methodology

4.1 Our system

Figure 1 illustrates how our system operates. In our initial approach, we collect video data from social media platforms and perform several processing operations to prepare the data. The data is then produced to the server via Kafka. The server proceeds to process and use the experimentally validated models along with PySpark to make predictions and generate final results.

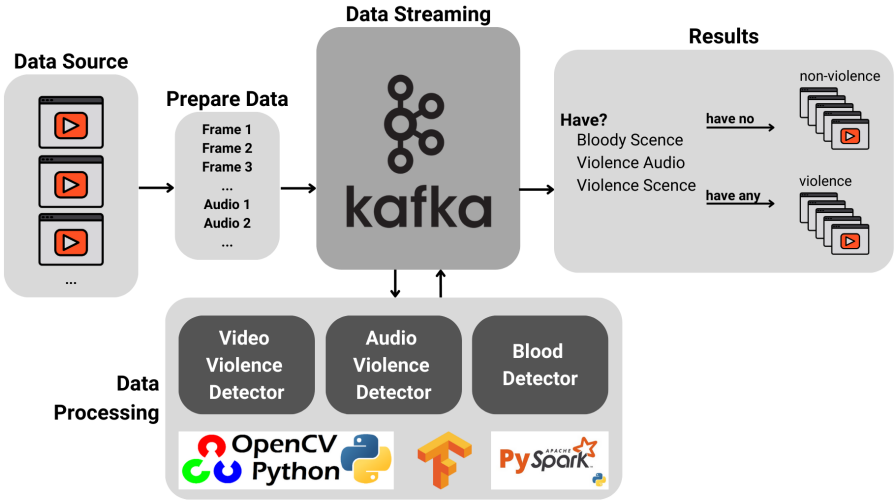

Fig. 1: Our real-time violence and hate speech detection in video system

Input: Video clips, which need to be detected whether they are safe for children or not.

Output: Label Violence or Non-Violence. Explained in detail in the section 4.2

4.2 Training for Violence Multi-class in Video

In addition to training models from pre-labeled datasets containing Violence and Non-Violence categories, we also employ a pre-trained model from Shakthi-Dhar [15] to support the task of detecting bloody features, and the AssemblyAI API [16] for detecting hate speech from audio features. We stipulate that if a video does not contain elements of fighting, violence, blood, or hate speech, it is considered safe for children.

Feature Extraction

- **Motion Features:** To perform feature extraction, we utilized the significantly pre-trained VGG19 model [19]. Compared to using self-extracted features from the model or employing a CNN network, we found that extraction using pre-trained models on the same task (violence or non-violence) optimized execution time and improved performance. The input to these networks is a sequence of video capture frames.
- **Blood Features:** Similar to motion features, blood detection employs ResNet10 [20] for feature extraction, demonstrating clear performance improvements. However, unlike action scenes involving fighting, blood detection relies on color and may only constitute a very small part of the frame. Therefore, we

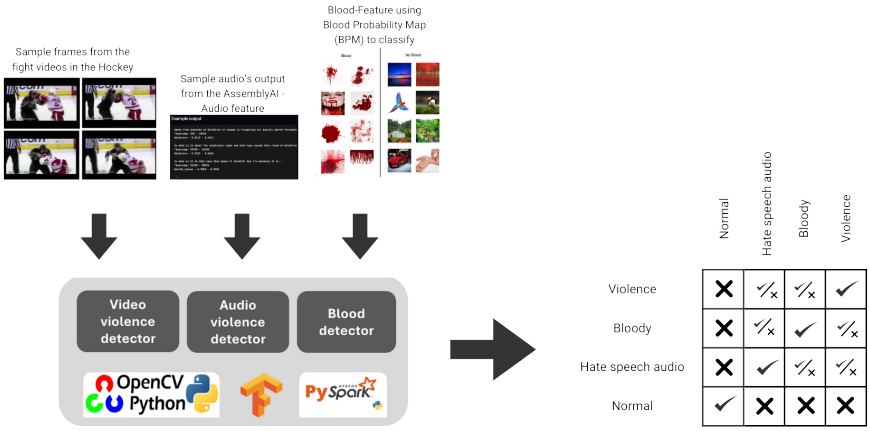

Fig. 2: The overview of our Violence Mutli-class system

also utilize a Blood Probability Map (BPM) [21]. This map has the same size as the input image and contains the blood probability values for every pixel.

- **Audio Features:** Using the AssemblyAI API, audio is transcribed into text and then classified to determine the presence of hate speech. This AI application is relatively high-performing and efficient, with execution time nearly parallel to the video.

Feature Classifcation In this study, we employ three different models, each utilizing distinct feature extraction techniques. After training the models with the extracted features, the output is their probabilities in various classes. These probability classes are combined and transformed back into a binary classification problem of Violence or Non-Violence. Using three models simultaneously impacts runtime; however, we have optimized the execution time as much as possible, prioritizing the accurate detection of inappropriate content for children.

4.3 Integration of Apache Kafka and Apache Spark

Apache Kafka The integration of Kafka in the process of managing and approving videos on social media platforms increasingly demonstrates its importance and potential in enabling timely review and filtering of harmful content unsuitable for children engaging in social media. This section discusses the benefits of utilizing Kafka and the data transmission from users to servers.

In our system, Kafka serves as the central data ingestion hub, receiving video data from users for server processing. Videos from various sources are efficiently recorded and transmitted to the server via Kafka Producer. The integration of Kafka into social media video management and approval offers the following

advantages: **(1) High-throughput data transmission**. Videos from multiple sources generate an enormous volume of real-time data, especially for social networks with massive user bases. The system's parallelization and data distribution capabilities ensure smooth processing of video streams from multiple sources, minimizing congestion and increasing the speed of video review to detect harmful content promptly and accurately. **(2) Fault tolerance and reliability**. Continuous monitoring is crucial in the video approval process, and system reliability is paramount. Kafka's fault-tolerant design ensures that even in the event of data transmission disruptions, data integrity is maintained. The use of replication mechanisms and distributed storage ensures that no data is lost, contributing to the development of a robust approval system. **(3) Real-time data ingestion capability** makes Kafka an ideal choice for easily collecting video segments from various sources. Kafka ensures continuous data flow, allowing the system to process videos requiring approval in a timely manner.

To implement Kafka for our problem, we consider each social media account as a Kafka Producer and our server as a Kafka Consumer.

Apache Spark Apache Spark is a powerful data processing platform capable of handling large datasets with high performance. Implementing Spark in the system brings several significant benefits: **(1) Rapid and efficient data processing** due to parallel processing capabilities and performance optimization. This is crucial for managing the large volume of data from videos, especially when processing within a short timeframe is necessary to detect and remove inappropriate content promptly. (2) With Spark Streaming, the system can **process video data streams in real-time**, allowing for quick detection and immediate prevention of unsuitable content, thus protecting children from unhealthy material. **(3) Spark can easily scale** to handle larger data volumes as the system grows, ensuring it meets the increasing demands in processing and analyzing harmful social media content in the future. **(4) Easy integration with other technologies and high flexibility** are among Spark's outstanding advantages. It enables the construction of a flexible and efficient video censorship system, facilitating the development of additional analytical tools, particularly Kafka, to enhance the system's capabilities.

5 Experiments

In this study, we conducted experiments with different models on three tasks: detecting violent scenes, identifying bloody images, and detecting hate speech through audio. Our goal was to identify the models with the best performance and the optimal runtime.

5.1 Violence motion detection

To detect violent motion in videos, we employ various models, including CNN-Long Short-Term Memory [22], Conv3D [23], Convolutional LSTM [24], and our

proposed model: Hybrid CNN-LSTM-Conv3D, which combines the advantages of the three aforementioned models. The models are trained on a large dataset consisting of videos containing violent and non-violent behavior.Table

Table 2 describes the experimental results for the task of detecting violent actions. It can be seen that across all three datasets, the proposed Hybrid CNN-LSTM-Conv3D model shows superior accuracy compared to the other models. Specifically, the model achieved 98% accuracy on the Hockey Dataset, 100% on the Movies Dataset, and 99.5% on the Violent-Flows Dataset. Other models, such as CNN-LSTM, Conv3D, and ConvLSTM, had lower accuracy, ranging from 93.8% to 97.1%. The proposed model also had lower loss rates compared to the other models, indicating more precise prediction capabilities. The loss rates for this model are 0.20 on the Hockey Dataset, 0.18 on the Movies Dataset, and 0.22 on the Violent-Flows Dataset. Other models had higher loss rates, ranging from 0.25 to 0.40. One of the greatest advantages of the proposed model is its very high processing speed, reaching up to 40 frames per second (fps) on the Hockey Dataset, 45 fps on the Movies Dataset, and 34 fps on the Violent-Flows Dataset. This processing speed is higher than all the previously published models.

Models	**Datasets**	**Accuracy**	**Cross Entropy Loss**	**Runtime (Hz)**
CNN-LSTM	Hockey Dataset	97.1	0.25	31Hz
Conv3D	Hockey Dataset	95.5	0.35	27Hz
ConvLSTM	Hockey Dataset	96.5	0.30	29Hz
Proposed	Hockey Dataset	**98.0**	**0.20**	**30Hz**
CNN-LSTM	Movies Dataset	95.3	0.32	30Hz
Conv3D	Movies Dataset	93.8	0.40	25Hz
ConvLSTM	Movies Dataset	94.7	0.35	28Hz
Proposed	Movies Dataset	**100**	**0.18**	**32Hz**
CNN-LSTM	RLVS Dataset	96	0.28	32Hz
Conv3D	RLVS Dataset	94,2	0.38	26Hz
ConvLSTM	RLVS Dataset	95.1	0.33	29Hz
Proposed	RLVS Dataset	**99.5**	**0.22**	**28Hz**

Table 2: Experimental results table on the task of violence motion detection

5.2 Bloody detection

Table 3 presents the results of the models on the task of detecting blood in video images. This table shows that the MobileNetV2 [25] model has impressive performance with an accuracy of 95%, Precision of 94%, Recall of 96%, and an F1-score of 95%. The training time for this model is 10 hours, reasonable and not excessively long compared to other models such as ResNet50 (12 hours), VGG16 (8 hours), and InceptionV3 (14 hours). The MobileNetV2 model not only has high accuracy but also a faster training time than some other well-known models, making it a suitable choice for the current task.

Models	Accuracy	Precision	Recall	F1-score	Training time (hour)
ResNet50	93	92	94	93	12
VGG16	90	89	91	93	**8**
InceptionV3	92	91	93	92	14
EfficienNetB0	94	93	95	94	9
MobileNetV2	**95**	**94**	**96**	**95**	10

Table 3: Experimental results on the task of blood detection

5.3 Violence audio detection

We use the speech recognition and content moderation features from Assembly AI to identify violent audio from the input video. This process includes analyzing audio characteristics, detecting hate speech, and consolidating harmful labels into an "unsafe" label for the entire audio portion of the video. These techniques help increase accuracy in detecting violent content and hate speech in videos while minimizing false detection rates.

5.4 Result

Fig. 3: Example of violence detected in video

To achieve the goal of this task, after obtaining results from the three tasks of detecting violent scenes, detecting blood, and detecting hate speech, our system will label the video as "safe for children" if it meets all three criteria of "no violent actions," "no blood scenes," and "no hate speech". We have an example of our results on a video at Fig 3. Testing on 50 short videos collected from YouTube, which we self-assessed for child-appropriateness, we achieved 90% accuracy and 85% recall, demonstrating good predictive capability for identifying videos unsuitable for children.

6 Conclusions

In this study, we proposed a system for managing and detecting videos containing content inappropriate for children. Our proposed system utilizes Assembly AI, MobileNetV2, and a Hybrid CNN-LSTM-Conv3D model. The system shows strong potential when integrating Kafka and Spark in the detection and elimination of harmful videos for children.

References

1. Children's YouTube is still churning out blood, suicide and cannibalism, https://www.wired.com/story/youtube-for-kids-videos-problems-algorithm-recommend/, last accessed 2024/07/23.
2. YouTube Kids app is STILL showing disturbing videos, http://www.dailymail.co.uk/sciencetech/article-5358365/YouTube-Kids-app-showing-disturbing-videos.html, last accessedvtv 2024/07/23.
3. TikTok, https://www.statista.com/topics/6077/tiktok/, last accessed 2024/07/23.
4. Tr em Vit Nam s dng mng xã hi ngày càng nhiu, https://vtv.vn/news-20230202120512126.htm, last accessed 2024/07/23.
5. Machaca Arceda, V. et al.: Real time violence detection in video. Presented at the January 1 (2016). https://doi.org/10.1049/IC.2016.0030.
6. Zhao, Y. et al.: Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors. 2018, (2018). https://doi.org/10.1155/2018/7316954.
7. Abdali, A.-M.R., Al-Tuma, R.F.: Robust Real-Time Violence Detection in Video Using CNN And LSTM. Presented at the March 27 (2019). https://doi.org/10.1109/SCCS.2019.8852616.
8. Nam, J. et al.: Audio-visual content-based violent scene characterization. Presented at the October 4 (1998). https://doi.org/10.1109/ICIP.1998.723496.
9. Weakly-Supervised Violence Detection in Movies with Audio and Video Based Co-training.
10. Datta, A. et al.: Person-on-person violence detection in video data. Presented at the August 11 (2002). https://doi.org/10.1109/ICPR.2002.1044748.
11. Deniz, O. et al.: Fast violence detection in video. Presented at the January 1 (2014).
12. Violence Detection in Videos, last accessed 2021/09/18.
13. Soliman, M.M. et al.: Violence Recognition from Videos using Deep Learning Techniques. Presented at the December 1 (2019). https://doi.org/10.1109/ICICIS46948.2019.9014714.
14. Violence detection in video using computer vision techniques, last accessed 2011/08/29.
15. Shakthi-Dhar/ML_Model-FaceBloodIdentifier, https://github.com/Shakthi-Dhar/ML_Model-FaceBloodIdentifier, last accessed 2024/07/23.
16. Identifying hate speech in audio or video files | AssemblyAI Docs, https://www.assemblyai.com/docs/guides/identifying-hate-speech-in-audio-or-video-files, last accessed 2024/07/23.
17. Bin Jahlan, H.M., Elrefaei, L.A.: Detecting Violence in Video Based on Deep Features Fusion Technique. abs/2204.07443, (2022). https://doi.org/10.48550/arXiv.2204.07443.

18. Pang, W.-F. et al.: Violence Detection in Videos Based on Fusing Visual and Audio Information. Presented at the June 6 (2021). https://doi.org/10.1109/ICASSP39728.2021.9413686.
19. Bansal, M. et al.: Transfer learning for image classification using VGG19: Caltech-101 image data set. (2021). https://doi.org/10.1007/S12652-021-03488-Z.
20. Gong, J. et al.: ResNet10: A lightweight residual network for remote sensing image classification. Presented at the January 1 (2022). https://doi.org/10.1109/icmtma54903.2022.00197.
21. Jones, M.J., Rehg, J.M.:Statistical color models with application to skin detection. In: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), vol. 1, pp. 274–2801 (1999).
22. Agga, Ali, Ahmed Abbou, Moussa Labbadi, Yassine El Houm, and Imane Hammou Ou Ali. "CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production." Electric Power Systems Research 208 (2022): 107908.
23. Park, Jae-Hyuk, Mohamed Mahmoud, and Hyun-Soo Kang. "Conv3D-based video violence detection network using optical flow and RGB data." Sensors 24, no. 2 (2024): 317.
24. Shi, Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." Advances in neural information processing systems 28 (2015).
25. Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520. 2018.