2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE 2024)
25-27 April, 2024, Gazipur, Bangladesh

Generalising Violence Detection with a New Near-Real-World Violence Dataset

Mahmudul Haque[†], Hussain Nyeem[§], Tareque Bashar Ovi[§], Al Nahid[§], Md. Sabbir Hossain Molla[§], Md. Tanjim Mahmud Tuhin[§], Fardin Shahab[§], Ayat Subah Alam[§] and Saadia Binte Alam[†‖*]

[†]Department of Computer Science and Engineering, Independent University, Bangladesh, Dhaka 1229, Bangladesh.
[§]Department of Electrical, Electronic and Communication Engineering (EECE),
Military Institute of Science and Technology (MIST), Mirpur Cantonment, Dhaka–1216, Bangladesh
[‖]Center for Computational & Data Sciences, Independent University, Bangladesh, Dhaka 1229, Bangladesh

Emails: mahmud.eece@gmail.com, h.nyeem@eece.mist.ac.bd, ovitareque@gmail.com,
al2453721@gmail.com, sabbir.eece@gmail.com, tanjimtuhin06@gmail.com, fardinshahab5306@gmail.com,
ayatsalam.uni@gmail.com, [*]saadiabinte@iub.edu.bd
[*]Corresponding Author(s)

***Abstract*—Detecting and classifying violence is crucial for public safety and addressing societal violence. DL models have greatly improved the automation of violence detection systems by effectively capturing intricate visual patterns. However, the quality and diversity of the training data greatly impact the effectiveness of these models. Existing datasets may be biased towards specific situations, limiting their practical use. To address this limitation, we introduce the Movie Clip (MC) dataset to enhance the generalisability of Automated Violence Detection and Classification (AVDC) systems. The MC dataset encompasses a broad spectrum of near-real-world violent actions, incorporating diverse demographics, environmental circumstances, and cultural elements extracted from movies. Consequently, it accurately reflects the complexity and diversity of real-world violent scenarios. The potential of the new dataset is investigated against its existing counterparts, like Hockey Fight (HF) and AIRTLab datasets. These datasets are used to train the ConvLSTM models based on the VGG16 and VGG19 architectures. The proposed dataset significantly improves AVDC model generalisation, outperforming the generalisability of current datasets, thereby advancing violence detection and facilitating the development of more robust and efficient AVDC systems.**

***Index Terms*—movie clip, violence detection, dataset, ConvLSTM, CNN, LSTM**

I. Introduction

Violence is a significant problem in modern society, posing threats to well-being, public safety, and social cohesion [1], [2]. It manifests in various forms, such as street crimes, domestic violence, cyberbullying, and extremism. Detecting and preventing violence is crucial for governments, law enforcement, and communities worldwide. However, continuous monitoring is challenging due to the potential for human error. Thus, the development of a vision-based Automated Violence Detection and Classification (AVDC) system holds great importance [3].

For the AVDC systems, deep learning (DL) based pattern recognition has demonstrated great promise. DL, a subfield of machine learning, has revolutionised computer vision through the utilisation of deep neural networks (DNNs) to learn hierarchical representations directly from input data [4]. This paradigm shift has led to significant progress in various computer vision tasks, enhancing accuracy, efficiency, and versatility. Convolutional neural networks (CNNs) and end-to-end learning have played significant roles in transforming tasks such as image classification, object detection, semantic segmentation, and generative models [5].

Characterizing human behaviour for violent activities has always been challenging in computer vision research. Despite notable advancements, there are still key limitations in this field, including optical flow discontinuities [6], camera aperture problems, and challenges related to illumination and feature tracking initialisation [7]. To address these challenges, LSTM-based dual-stream network [8] employed a late-fusion approach, combining appearance, motion, and audio features, and achieved state-of-the-art performance in that year. Later, the classic action recognition of temporal segment was utilised in the FightNet model [9] to detect complex visual violence interactions. Some computationally efficient approach were also taken such as Vijeikis *et al*. [10] utilised MobileNetV2 along with LSTM network for violence detection by extracting temporal features.

Subsequently, ConvNet, a video content understanding approach that considers spatiotemporal features, has been widely used in violent video detection [11], [12]. Although these models have demonstrated promising results, the effectiveness of these DL models heavily relies on the quality and diversity of the training datasets. Therefore, it is crucial to train models with datasets that can help the models learn features that are general for the actions associated with violent scenes which make the trained model capable of detecting a wider range of violent scenes

This paper presents a new dataset and justifies its effectiveness in improving the generalisability of AVDC systems in detecting a wide range of violent activities. Furthermore, the

979-8-3503-8828-2/24/$31.00 ©2024 IEEE

potential of this dataset is investigated in comparison to its existing counterpart. Existing AVDC datasets, such as Hockey Fight (HF) [13] and AIRTLab [14], suffer from significant bias towards specific contexts and situations. Consequently, models trained on these datasets may struggle to generalize to unseen instances of violence, posing challenges for the practical implementation of DL model-based violence detection systems.

To address the critical need for a diverse AVDC dataset, we have developed a Movie Clip (MC) dataset with a comprehensive collection of visual data samples extracted from movie clips (Sec. II). Our dataset incorporates a wide range of near-real-world violent activities, aiming to overcome the limitations of existing datasets (Sec. III). By ensuring diversity across various dimensions, such as the types of violence, environmental contexts, demographics, and cultural factors, the new dataset demonstrates greater effectiveness for the DL-based AVDC systems capturing the complex and heterogenous real-world violent activities (Sec. IV).

II. Development of a New Dataset

In this section, we will evaluate the strengths and limitations of the current AVDC datasets. Our analysis will primarily concentrate on two extensively utilised datasets: HF [13] and AIRTLab [14]. Furthermore, we will introduce a new dataset that has been developed to address the limitations of the existing datasets. Throughout our assessment, we will specifically emphasize the fundamental attributes of the sampled video frames, volumes, scene diversity, and other technical considerations that are critical for the development of a comprehensive dataset.

A. Existing Datasets

1) HF Dataset: The HF dataset, developed by Nievas *et al*. [13], contains 1,000 clips from National Hockey League (NHL) games, representing both violent and non-violent actions. It is a valuable resource for AVDC tasks. Each clip consists of 41 to 50 frames with a resolution of 720×576 pixels. The dataset has undergone manual annotation to classify the clips into two categories: *violent* and *non-fight*. The dataset includes 500 clips depicting violent scenes, specifically fighting scenes in NHL games, and 500 clips depicting non-violent scenes, primarily showcasing general gameplay. Fig. 1a provides samples extracted from the HF dataset. Notably, all videos in this dataset have a standardised duration of approximately two seconds and maintain consistent frame sizes. They also exhibit similar backgrounds and background motions, ensuring a coherent visual context throughout the entire dataset.

2) AIRTLab dataset: The AIRTLab dataset [14] was created specifically to evaluate the robustness of AVDC models in the presence of false positives in non-violent video clips with rapid movements [15]. It consists of 350 video files with an average duration of 5.63 seconds. The videos have a resolution of 1920×1080 pixels and a frame rate of 30 frames per second, using the H.264 codec. The dataset is organised into two directories: 'non-violent' and 'violent' with different camera perspectives capturing the same activities. There are 120 instances of non-violent behaviour, and 230 instances of violent behaviour. These scenes were primarily performed by actors to simulate both violent and non-violent interactions. To create the dataset, actors were recorded using two cameras, resulting in 2–4 clips for each violent and non-violent action. Initially, the actors performed various violent actions such as kicking, punching, slapping, beating, and simulated gunshots. Subsequently, non-violent actions like high-fives and hugging were enacted. Fig. 1b displays sample clips from the AIRTLab dataset.

B. Scope of Development

To develop a new dataset for AVDC models, it is important to consider several factors. Firstly, a diverse representation of violent scenes is crucial, encompassing various types of violent behaviours, environments, and contexts. This diversity enables the model to learn robust features and generalize effectively to real-world scenarios. Secondly, the dataset should include an ample number of non-violent scenes to achieve a balanced training set and prevent bias towards violence. This ensures the model can effectively discriminate between violent and non-violent behaviours.

Furthermore, manual annotation of the dataset is essential to provide accurate and consistent ground truth labels for training and evaluation. The annotation process should be meticulous to ensure reliable evaluation metrics. Additionally, factors such as video quality, resolution, and frame rate need to be considered to ensure compatibility with different surveillance systems and video sources.

However, existing AVDC datasets do not encompass general application scenarios. As discussed above in this section, the HF and AIRTLab datasets address specific features of violent scenes but are insufficient for developing a generalised AVDC system. To overcome this limitation of the existing dataset, we aimed to develop a new dataset containing a wide range of violent scenes to facilitate the development of a generalised AVDC system suitable for real-world violence detection.

C. The Proposed MC Dataset

We have developed an extensive dataset consisting of 1,377 video clips extracted from various movies, wherein near real-world situations are portrayed. Each clip in the dataset adheres to a standardised resolution of 1920×1080 pixels and maintains a frame rate of 24 frames per second. The duration of these clips varies, ranging between 97 and 172 frames. To ensure clarity and organisation, we have divided the dataset into two distinct categories: violent and non-violent. Clips depicting violence are labelled and included in the violent category, while those without violence are assigned to the non-violent category. Fig. 1c displays selected samples from the dataset, providing a visual representation.

Our dataset comprises a total of 913 violent and 464 non-violent movie clips. For a comprehensive overview, we present the details of all three datasets in Table I, which demonstrates the novelty of our approach to creating

the proposed dataset. The MC dataset can be accessed at https://figshare.com/articles/dataset/23643555.

III. Experiment Settings

In this section, we outline the experimental settings used to assess and analyse the performance of DL-based AVDC models. Our evaluation focuses on the proposed MC dataset, as well as the HF [13] and the AIRTLab [14] datasets. We provide technical details regarding the preprocessing of the datasets, the architecture of the ConvLSTM-based AVDC models, their hyperparameters, and the method used to measure the diversity of the MC dataset against the HF and AIRTLab datasets.

A. Dataset Pre-Processing

To ensure uniformity and facilitate analysis, we adjusted the resolution of the video clips in our datasets. The original clips exhibited varying resolutions and aspect ratios across the datasets, necessitating even resizing while preserving the format. We applied appropriate padding techniques to ensure that all clips maintained a consistent aspect ratio of 1:1. After preprocessing, the dataset was divided into three distinct sets: training, validation, and testing datasets.

For the model development phase, 70% of the clips were allocated for training the AVDC models, while 15% were reserved for validation purposes. The remaining 15% of the video clips were exclusively used to assess the performance and generalisation abilities of the trained models. We aimed to minimize biases, achieve a balanced representation of violent and non-violent scenes, and adhere to conventional practices for training and testing AVDC models. These procedures were employed during the dataset preparation and processing stages to ensure adherence to established standards.

B. AVDC Models

In this study, two DL models have been chosen to demonstrate the performance of our dataset. These models are transfer learning-based variations of ConvLSTM models [15]. ConvLSTM is a variant of LSTM that incorporates convolutional processes while transitioning between states. Multiple ConvLSTMs can be used to create an encoding-prediction framework that is effective in obtaining spatiotemporal properties.

Traditionally, the input data for LSTM is one-dimensional; spatial sequence data like video, satellite, and radar image datasets are not appropriate. Hence, ConvLSTM was designed with 3D input data where both spatial and temporal features are considered. As a result of the deployment of the convolutional framework, all inputs X_t, cell states C_t, hidden states H_t, and gates (I_t, F_t, O_t) are three-dimensional tensors $\in \mathbf{R}^{\mathbf{b}\times\mathbf{b}\times\mathbf{k}}$, the initial two dimensions capture spatial characteristics, whereas the final dimension acquires spectral depiction of features. subsequently the temporal dependencies were developed via generating time series for H and C as illustrated in Fig.2. The input X_t and past states C_{t-1}, H_{t-1} are inputs to the ConvLSTM for predicting the future states C_t, H_t. Equation (1) summarizes the essential equations of the ConvLSTM according to [16].

$$
\begin{aligned}
I_t &= \sigma\left(W_{XZ} * X_t + W_{HI} * H_{t-1} + W_{CI} \circ C_{t-1} + b_I\right) \\
F_t &= \sigma\left(W_{XF} * X_t + W_{HF} * H_{t-1} + W_{CF} \circ C_{t-1} + b_F\right) \\
C_t &= F_t \circ C_{t-1} + I_t \circ \tanh\left(W_{XC} * X_t + W_{HC} * H_{t-1} + b_C\right) \\
O_t &= \sigma\left(W_{XO} * X_t + W_{HO} * H_{t-1} + W_{CO} \circ C_t + b_0\right) \\
H_t &= O_t \circ \tanh\left(C_t\right)
\end{aligned}
\tag{1}
$$

Where $W, b, \sigma, *$ and $\circ$ represents coefficient matrix, bias vector, sigmoid function, convolution operation and Hadamard product respectively.

For initial feature extraction, we have separately considered VGG-16 and VGG-19 [17]-based convolutional backbones. VGG-16 and VGG-19 are CNN architectures with 16 and 19 convolutional layers, respectively, which help the model understand the spatial features of the data. These spatial features are then used to train the LSTM layers to recognise the temporal pattern across these spatial features and enable the recognition of violent scenes with appropriate hyper-parameters.

C. Hyper-parameter

To ensure robust training, we employed 100 epochs for both the VGG-16 and VGG-19-based ConvLSTM models. Each epoch represents a complete iteration through the entire dataset. We used a batch size of 24, which facilitated efficient parallel processing and gradient updates during training. These choices strike a balance between computational efficiency and effective model learning.

For the output layer activation, we utilised the sigmoid function. This activation function is well-suited for binary classification tasks, as it enables effective discrimination between violent and non-violent classes. By employing the sigmoid activation, we encouraged the models to output a probability value that indicates the likelihood of a violent scene.

To optimize the models' performance during training, we employed the Adam optimizer, known for its effectiveness in optimizing deep neural networks. We initialised the learning rate as 1e-5, allowing the model to gradually update its parameters to minimize the loss function and converge to an optimal solution.

To measure the discrepancy between predicted and actual labels, we utilised the binary cross-entropy loss function. This loss function is commonly used in binary classification scenarios, aiding the models in learning and adjusting their parameters accordingly. By minimizing the binary cross-entropy, we encouraged the models to make accurate predictions and classify violent and non-violent scenes more effectively.

Furthermore, we note the parameter counts for our models. The VGG-16 + ConvLSTM model comprises a total of 19,598,401 parameters, while the VGG-19 + ConvLSTM model has 20,024,384 parameters. These parameters capture the models' weights and biases, enabling them to learn and extract relevant features from the input data, thus enhancing their prediction accuracy.

TABLE I: Overview of the HF, AIRTLab, and MC datasets.

Criteria	**HF dataset**	**AIRTLab**	**MC dataset**
Resolution	720×576	1920×1080	1920×1080
Size (clips)	1000	350	1377
FPS	41-50	30	24
Violent scenes	500	230	913
Non-violent scenes	500	120	464
Duration (approx)	2 sec	5.63 sec	5-7 sec
Source	NHL games	Staged acting	Movies
Considerations	Fighting in Sports	Dummy fighting Weapon props	Fighting, Aggression, Blood, Weapons, Murder and other physical altercations

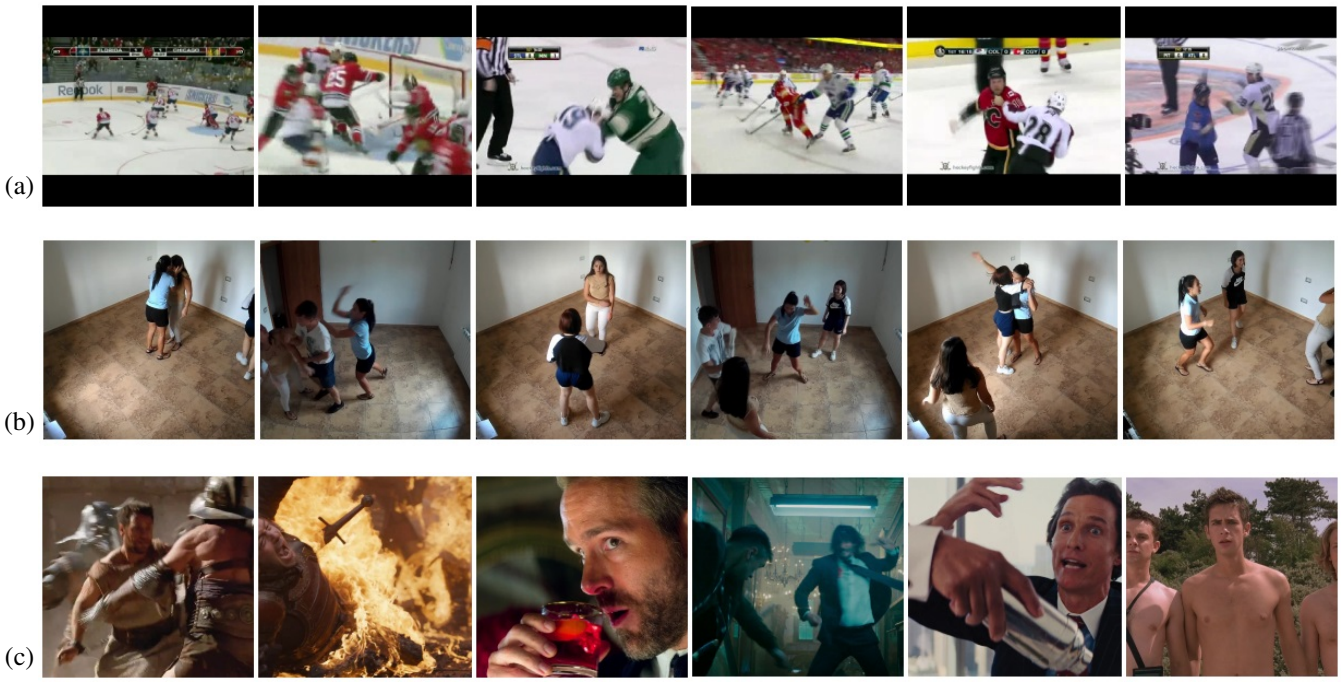

Fig. 1: Sample frames of different datasets: (a) HF, (b) AIRTLab, and (c) MC (*proposed*).

D. Evaluation Metrics

The performance evaluation of the AVDC system to determine the effect of the proposed MC dataset in improving its generalizability involved training and validating the AVDC models on three distinct datasets: MC, AIRTLab, and HF. To identify the optimal models for AVDC, we considered the trade-off between training and validation loss values. Subsequently, each of these AVDC models underwent evaluation using separate test datasets from all three aforementioned datasets. The schematic representation of this process is depicted in Fig. 3. The subsequent sections present and assess the performance of the models on each test dataset, using the F1 score and area under the curve (AUC) as evaluation metrics.

However, while these metrics aid in understanding a model's performance on a specific dataset, they do not capture its ability to generalize across different datasets. To address this, we introduce a cross-dataset diversity factor (δ_X), defined by Equation (2), where α denotes the accuracy of the model on the test dataset derived from the same dataset as the training dataset (but not used in training), and α' represents the accuracy of the model on the test dataset obtained from other datasets.

$$\delta_X = \frac{\alpha - \alpha'}{\alpha} \tag{2}$$

Using this evaluation metric, the deviation in test accuracy of the model for each dataset can be determined. A higher value indicates a lower generalisation of the model, and vice versa. In essence, this metric evaluates the generalisation achieved by the AVDC model due to the diversity of features that represent violent actions in a given training dataset.

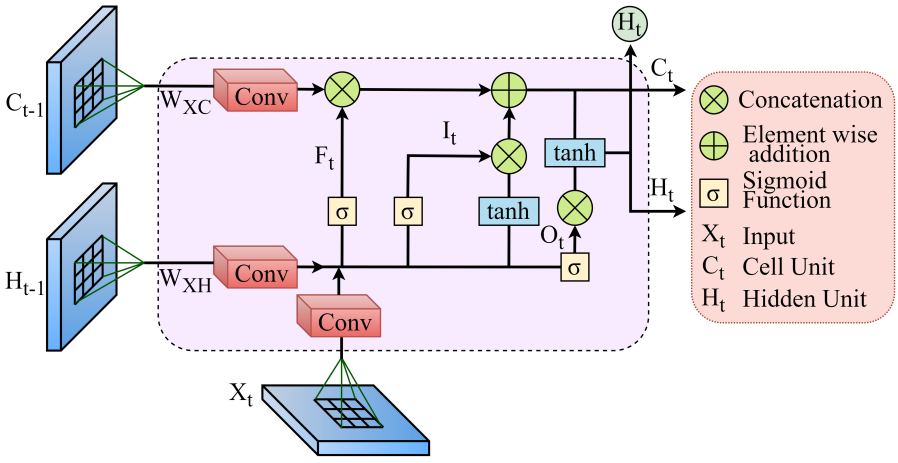

Fig. 2: Internal Architecture of ConvLSTM

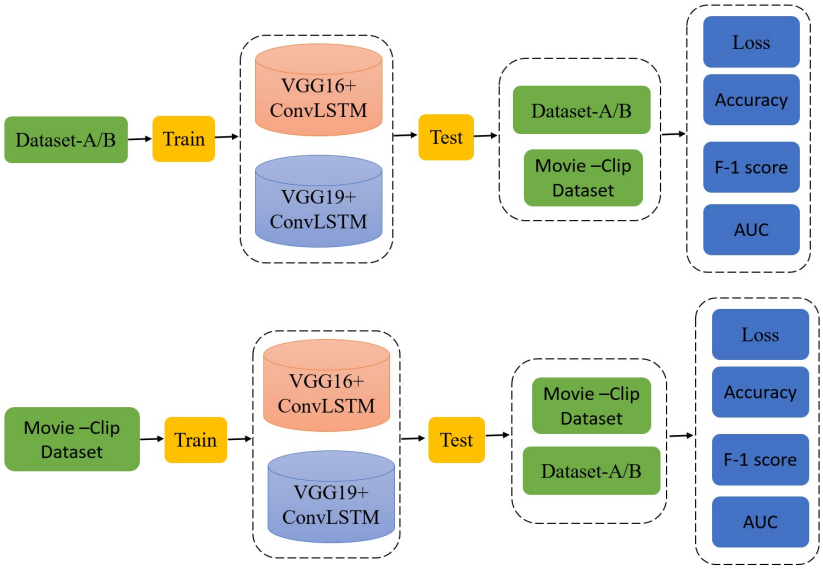

Fig. 3: Process of Model Evaluation

IV. Result and Analysis

In this section, we investigate the potential of our proposed MC dataset to improve the generalisation of ConvLSTM models based on VGG16 and VGG19 for violence detection in reference to the HF [14] and AIRTLab datasets. We analyse the performance of the models in terms of accuracy, F1 score, and AUC. Furthermore, we determine the respective δ_X values based on the accuracies of the models on different test datasets.

Table II presents the performance of different variants of the AVDC approach using VGG16+ConvLSTM, trained and tested with MC, HF, and AIRTLab datasets. The impact of the MC dataset on the AVDC model's performance, measured by δ_X, is significantly better compared to the HF and AIRTLab datasets. The overall δ_X values for the models trained with the MC, HF, and AIRTLab datasets are 0.2514, 0.3645, and 0.3568, respectively. Therefore, the AVDC model trained with our proposed MC dataset achieves 29.54% and 31.03% higher generalisation in terms of handling diversity compared to the state-of-the-art HF and AIRTLab datasets.

Similarly, Table III illustrates the performance of different variants of the AVDC approach using VGG19+ConvLSTM, trained and tested with MC, HF, and AIRTLab datasets. The δ_X values obtained for the models are 0.3497, 0.6696, and 0.4647, respectively. These results indicate that the model trained with our MC dataset achieves an enhanced generalisation of 47.77% and 24.75% compared to the HF and AIRTLab datasets, respectively.

Overall, our findings demonstrate that the proposed MC dataset significantly improves the generalisation capabilities of VGG16 and VGG19-based ConvLSTM models. These results provide strong evidence of the effectiveness and novelty of

TABLE II: Performance of VGG16-ConvLSTM

Test Dataset	Metric	Train Dataset		
		MC	**HF**	**AIRTLab**
MC	**Accuracy**	0.7439	0.4589	0.6618
	F1-Score	0.814	0.4717	0.7784
	AUC	0.6947	0.5039	0.5489
	δ_X	-	0.5118	**0.2537**
HF	**Accuracy**	0.4533	0.94	0.46
	F1-Score	0.6238	0.9396	0.5759
	AUC	0.4533	0.94	0.4599
	δ_X	**0.3906**	-	0.4813
AIRTLab	**Accuracy**	0.6604	0.7358	0.8868
	F1-Score	0.7954	0.8333	0.9143
	AUC	0.5	0.6111	0.8738
	δ_X	**0.1122**	0.2172	-

TABLE III: Performance of VGG19-ConvLSTM

Test Dataset	Metric	Train Dataset		
		MC	**HF**	**AIRTLab**
MC	**Accuracy**	0.7391	0.4541	0.3961
	F1-Score	0.8111	0.4593	0.3386
	AUC	0.6876	0.5038	0.4739
	δ_X	-	**0.5169**	0.5533
HF	**Accuracy**	0.433	0.94	0.5533
	F1-Score	0.6009	0.9412	0.2117
	AUC	0.4333	0.94	0.5533
	δ_X	0.4142	-	**0.3761**
AIRTLab	**Accuracy**	0.5283	0.566	0.8868
	F1-Score	0.5762	0.6933	0.9142
	AUC	0.5484	0.4825	0.8738
	δ_X	**0.2852**	0.3979	-

our work in creating a diverse dataset specifically designed for AVDC systems. Our MC dataset outperforms the state-of-the-art HF and AIRTLab datasets in terms of model generalisation, validating the significance of our contribution.

V. Conclusion

In this paper, we have addressed a crucial and timely requirement for a diverse and comprehensive dataset for AVDC systems. We have introduced the MC dataset, which overcomes the limitations of existing datasets by capturing the complexity and heterogeneity of real-world violent scenarios. Our experiments have demonstrated that the MC dataset surpasses well-known datasets like HF and AIRTLab when training VGG16 and VGG19-based ConvLSTM models. The results also indicate that the MC dataset significantly improves the generalisation capabilities of AVDC models, exhibiting notable performance in terms of accuracy, F1 score, and AUC. The diversity of the MC dataset, encompassing various dimensions such as violence types, environmental contexts, demographics, and cultural factors, ensures that trained models can effectively detect a wide range of violent activities in near-real-world situations. Thus, this work introduces new possibilities for developing more reliable and effective violence detection systems by providing an unbiased dataset that encompasses a broader range of contexts and scenarios. The utilisation of DL techniques, with the power of the MC dataset, establishes the foundation for automated vision-based violence detection, thereby contributing to public safety and addressing the pressing issue of violence in our society.

References

[1] S. D. Hillis, J. A. Mercy, and J. R. Saul, "The enduring impact of violence against children," *Psychology, Health & Medicine*, vol. 22, no. 4, pp. 393–405, 2017.

[2] M. L. Capella, R. P. Hill, J. M. Rapp, and J. Kees, "The impact of violence against women in advertisements," *Journal of Advertising*, vol. 39, no. 4, pp. 37–52, 2010.

[3] A. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis, "Bidirectional convolutional lstm for the detection of violence in videos," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[4] M. L. Smith, L. N. Smith, and M. F. Hansen, "The quiet revolution in machine vision-a state-of-the-art survey paper, including historical review, perspectives, and future directions," *Computers in Industry*, vol. 130, p. 103472, 2021.

[5] M. Shah and R. Kapdi, "Object detection using deep neural networks," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2017, pp. 787–790.

[6] Efros, Berg, Mori, and Malik, "Recognizing action at a distance," in *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 2003, pp. 726–733.

[7] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1997, pp. 568–574.

[8] Q. Dai, R.-W. Zhao, Z. Wu, X. Wang, Z. Gu, W. Wu, and Y.-G. Jiang, "Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning." in *MediaEval*, vol. 1436, 2015.

[9] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violent interaction detection in video based on deep learning," in *Journal of physics: conference series*, vol. 844, no. 1. IOP Publishing, 2017, p. 012044.

[10] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveillance," *Sensors*, vol. 22, no. 6, p. 2216, 2022.

[11] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3d convolutional neural network," *Sensors*, vol. 19, no. 11, p. 2472, 2019.

[12] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3d convolutional neural networks," *IEEE Access*, vol. 7, pp. 39 172–39 179, 2019.

[13] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14*. Springer, 2011, pp. 332–339.

[14] M. Bianculli, N. Falcionelli, P. Sernani, S. Tomassini, P. Contardo, M. Lombardi, and A. F. Dragoni, "A dataset for automatic violence detection in videos," *Data in brief*, vol. 33, p. 106587, 2020.

[15] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, "Deep learning for automatic violence detection: Tests on the airtlab dataset," *IEEE Access*, vol. 9, pp. 160 580–160 595, 2021.

[16] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.