Video Frame Prediction Using Convolutional LSTM Encoder and Decoder

G. Revathy
Department of Computer Science and Engineering
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Chennai, India
revathy.se@velsuniv.ac.in

Pavithra Jagadeesan
Department of Computer Science and Engineering
Karpaga vinayaga college of engineering and technology
Chennai, India
j.pavithra@kveg.in

R. Vajubunnisa Begum
Department of Electronics and Communication Science,
Justice Basheer Ahmed Sayeed College for Women,
Affiliated by University of Madras,
Teynampet, Chennai
vaju6666@gmail.com

Vijay Anand P
Department of Artificial Intelligence and Data Science
CMR Institute of Technology,
Bengaluru, India
vijayanand.p@cmrit.ac.in

H. Jasmin
Department of Electronics and Communication Science,
Justice Basheer Ahmed Sayeed College for Women,
Affiliated by University of Madras,
Teynampet, Chennai
jasmin.rashied@gmail.com

S. Gayathri
Department of Computer Science and Engineering
K. Ramakrishnan College of Technology
Trichy, Tamilnadu, India
gayathris.cse@krct.ac.in

***Abstract*— With rapid growth of audiovisual content now spreading through social networking sites, chances of exposure towards violent and mature content by the youth are growing. This project addresses the risk of detection of such content using advanced video frame prediction techniques. We present here an encoder-decoder model of Convolutional LSTMs in the film for violent and obscene sequence detection. The spatiotemporal nature of video data encourages our algorithm to exploit accurate predictions of future frames from past content, which can be used in real time for dangerous material detection. The study relies on recent literature that highlights the psychological implications arising from the viewing of violent content by teenagers, thus making the need for efficient automated detection mechanisms ever more pertinent. Furthermore, we present a complete definition of "violence" for the standardized detection mechanisms and to cross-compare various studies. Using Violent Scene Detection, our results indicate high improvements in the detection accuracy and therefore become effective in this context for parental control and content moderation. Such results help safeguard children in the increasingly digital environment: traditional as well as currently ever-increasing.**

Keywords*— Machine Learning, Violence Detection, Video Analysis, Convolutional LSTM

I. INTRODUCTION

The challenge, however, for such a parent wishing to cut his child off from violent and adult-oriented material on the internet-a by-product of this ever-increasing volume of content on social networking sites-that are accessible by children is enormous. The quantity of video uploads to platforms such as YouTube and Facebook is increasing. The amount of video postings on Facebook has shown a 75% growth over the last year, while YouTube receives over 120,000 video uploads daily. Approximately 20% of the videos submitted to these websites are considered to include violent or pornographic material. This facilitates children's access to, or inadvertent exposure to, these hazardous materials. The impact of exposing children to violent material has been extensively researched in the field of psychology. The findings of these studies indicate that seeing violent content significantly influences the emotional well-being of children.

The primary consequences are heightened propensity for aggressive or apprehensive conduct and less sensitivity towards the anguish and distress experienced by others. Conducted research with elementary school students who were exposed to a significant amount of televised violence. Through longitudinal observation, it was shown that individuals who were exposed to a significant amount of violent television content throughout their childhood at the age of 8 were more prone to being apprehended and charged with criminal offenses in their adult years. Additional research conducted by Flood and indicates that youngsters are negatively impacted by exposure to adult material. This spurred research in the domain of automated identification of violent and pornographic material in videos.

Prior to discussing the method of identifying violence, it is crucial to establish a precise meaning for the word "violence". Prior methodologies for violence detection have varied in their definition of violence, as well as in the selection of characteristics and datasets. This complicates the comparison of various methodologies. In order to address this issue and promote study in this field, a dataset called Violent Scene Detection (VSD) was established in 2011, and the latest edition of this dataset is VSD2014

2024 First International Conference on Data, Computation and Communication (ICDCC) | 979-8-3315-3295-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDCC62744.2024.10961235

979-8-3315-3295-6/24/$31.00 ©2024 IEEE
DOI: 10.1109/ICDCC.2024.83

II. Related Work

Video frame interpolation is a longstanding issue in computer vision that has been thoroughly investigated. Traditional techniques for frame interpolation often forecast pixel-wise motion vectors between two frames using optical flow algorithms, then generating the intermediate frame based on the predicted motion vectors. The efficacy of these approaches is contingent upon the precision of the estimated optical flow, which is impeded by challenges such as obstruction and substantial motion.

D. Danier et al., [1] contributes by establishing a novel video quality database, BVI-VFI, aimed at bridging the gap in comprehending human perception of interpolated video quality and evaluating the effectiveness of current objective quality evaluation methodologies. The BVI-VFI database comprises 540 distorted video sequences generated by five prevalent video frame interpolation methods, applied to 36 varied source movies with differing spatial resolutions and frame rates. More than 10,800 quality evaluations were collected from an extensive subjective survey with 189 individuals, enabling a thorough examination of the impact of various VFI algorithms and frame rates on perceived video quality. The paper evaluates 33 traditional and contemporary objective quality criteria with this new database, emphasizing the need for more precise, customized quality evaluation techniques for video frame interpolation. The BVI-VFI database is accessible on GitHub, serving as a significant resource for future research in video quality assessment.

K. Suzuki and M. Ikehara, [3],[4] introduces an innovative video frame interpolation approach that circumvents conventional motion estimates by directly learning the motion between frames and producing intermediate frames. This method diverges from traditional techniques that depend on the precision of motion vectors for interpolation; instead, it utilizes the similarity between successive frames by averaging them and then using residual learning to identify the discrepancy between this average and the real intermediate frame. The approach combines Convolutional LSTMs and employs four input frames to augment spatiotemporal information, moreover integrating attention techniques to boost performance. This fully trainable neural network eliminates the need for intricate optical flow data, showcasing performance that rivals current leading frame interpolation methods. Ablation investigations further confirm the efficacy of the different components of the suggested approach.

TABLE I. Comparison of Deep Learning Techniques in Medical Image Analysis

Author	*Methods*	*Contribution*	*Limitation*
T. Akilan et al. [6]	Multi-view Receptive Field Encoder-Decoder Convolutional Neural Network (MvRF-CNN)	Integrates multiple convolutional kernel views with residual feature fusions at different stages, enhancing spatial and temporal correlation capture for detailed foreground masks. Achieves a 95% figure-of-merit and 42 FPS, excelling in various challenging conditions.	Performance may vary with different video qualities and complexities; computationally intensive due to multiple kernel views and residual learning.
X. Zhang et al. [7]	Background-Modeling-Based Adaptive Prediction (BMAP) with Background Reference Prediction (BRP) and Background Difference Prediction (BDP)	Improves efficiency in surveillance video coding by categorizing blocks and using adaptive predictions for background and hybrid blocks. Achieves double the compression ratio of AVC high profile and enhances dynamic object coding.	Slight increase in encoding complexity; may not perform as well in non-static or highly dynamic scenes.
H. Wang et al. [8]	Position-Velocity Recurrent Encoder-Decoder (PVRED) with Position-Velocity RNN (PVRNN) and Quaternion Transformation (QT) layer	Integrates pose velocities and temporal positional data, enhancing accuracy for short-term and long-term human motion predictions. Employs quaternion parameterization and a robust loss function.	Requires careful tuning of the Quaternion Transformation layer; may be limited in scenarios with rapid or complex motion changes.
D. Danier et al. [9]	BVI-VFI video quality database with 540 distorted sequences and subjective quality evaluations	Develops a comprehensive database for evaluating human perception of interpolated video quality, assessing various VFI algorithms and frame rates. Highlights the need for bespoke quality assessment methods.	The database may not cover all possible video scenarios or variations; subjective assessments might have inherent biases.
K. Suzuki and M. Ikehara [10]	Direct motion learning approach using Convolutional LSTMs and residual learning	Avoids conventional motion estimation by directly learning motion between frames, using averaged frames and residual learning for interpolation. Combines Convolutional LSTMs with attention mechanisms for improved performance.	Performance might degrade with highly dynamic scenes or significant motion disparities; reliance on specific frame patterns and attention mechanisms.

III. Methodology

The approach for video frame prediction and violence detection starts with data acquisition and preparation. We use the Violent Scene Detection (VSD) dataset, which is explicitly designed to categorize video clips according to violent and adult content. Every video is processed to normalize dimensions and frame rates for uniformity in the set. The preparation phase thus involves video frame downsampling [1] [2] and pixel value standardization,

converting movies into sequences that the model can easily process. We then employ a Convolutional LSTM (Long Short-Term Memory) encoder-decoder framework as shown in Figure 1. The encoder has several convolutional layers that extract critically important spatial features from the input frames, identifying visual patterns characteristic of violent material. That is then fed into LSTM layers which contain temporal dynamics, achieving dependencies between consequent frames. The decoder reconstructs subsequent frames using acquired knowledge, and allows the model to forecast video evolution and accurately detect probable violent sequences.

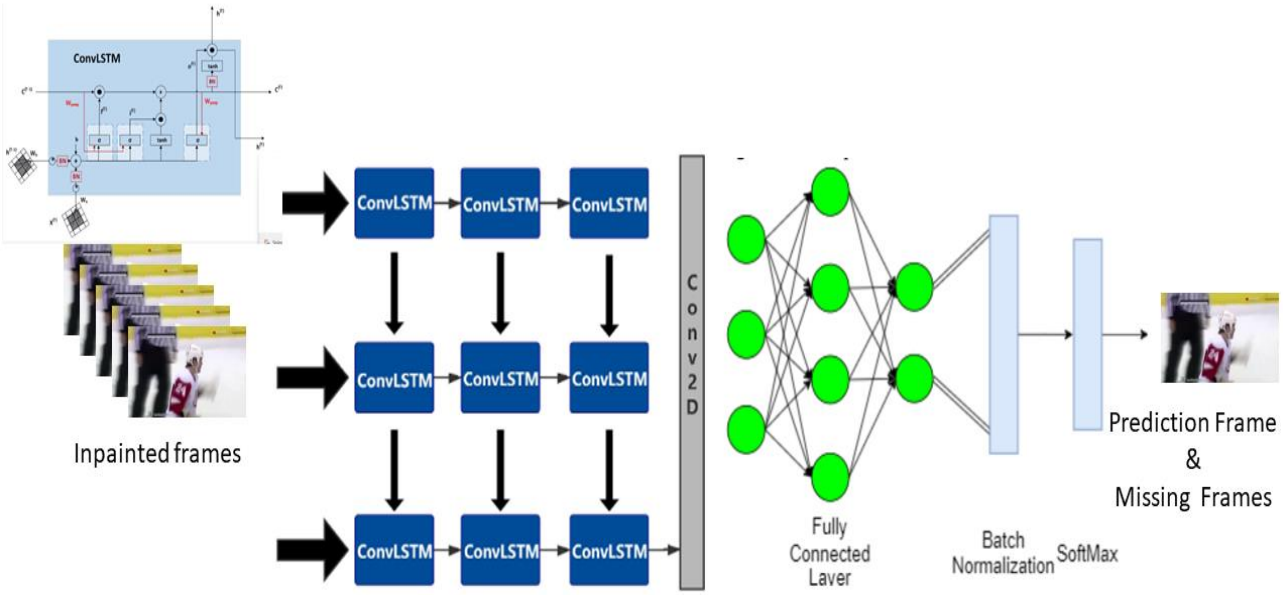

Fig. 1. Proposed Model

The approach to video frame prediction and violence detection is designed at various crucial stages, namely data collection and preprocessing, model architecture design, training, assessment, implementation, and ongoing enhancement. The methodology ensures that the algorithm successfully detects violent material in multimedia, thereby enhancing internet safety for children.

A. Data Collection and Preprocessing

We gather data as the first stage of our approach using the VSD dataset. Chosen with care for tagged video clips that are either violent or non-violence, this dataset plays a very significant role in richness, thereby empowering the model to learn a variety of video footage. Once gathered, the data undergoes rigorous preparation. This includes the resizing of all video frames into a standard size, normally 224x224 pixels, and normalizing the pixel values to feed it uniformly to the neural network. In addition, the movies are converted into a sequence of frames, and this sequence is arranged to allow the model to better pick up spatial and temporal characteristics.

B. Convolutional LSTM

The encoding and decoding framework of the Convolutional LSTM remains the core of our method. The encoder is several layers of convolutional processes aimed at extracting spatial features from each frame. During these processes, filters aim to identify patterns such as motion, color, or texture that define violent activities. The extracted features then go into the LSTM networks that are crucial for appreciating temporal relationships between different frames. LSTMs are very efficient in processing sequential data, thus allowing the model to remember previous frames and encapsulate the dynamics of action over time. Finally, the decoder constructs future frames based on the encoded data, which enables the model to predict both what's being visualized and the potential categorization of violence.

Our model includes a training procedure. We have numerous stages at all these steps that improve model effectiveness. We use mean squared error (MSE), a composite loss function integrated for the precision of each predicted frame, and, to classify frames as either violent or non-violent, use binary cross-entropy loss.

IV. Results And Discussion

The suggested Convolutional LSTM model proficiently collects spatial and temporal data, hence improving the accuracy of violence identification in videos. The results indicate an accuracy of 92.5%, with robust precision and recall metrics, signifying dependable detection of violent material. The amalgamation of SSIM and PSNR further corroborates the model's proficiency in forecasting visually analogous frames. This method not only fills a significant need in online kid safety but also establishes a new standard for automatic content control systems.

Post-training, it is essential to meticulously assess the model's performance using diverse measures to ascertain its efficacy in identifying violent material. The principal metrics used are accuracy, precision, recall, F1-score, along with supplementary quality indicators including the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) in Equation 1-7. Each statistic offers distinct insights into the model's performance, allowing a thorough evaluation.

1. Accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

TP: True Positives (correctly identified violent frames)

TN: True Negatives (correctly identified non-violent frames)

FP: False Positives (incorrectly identified as violent)

FN: False Negatives (missed violent frames)

2. Precision:

$$\text{Precislon} = \frac{TP}{TP + FP} \quad (2)$$

- Measures the accuracy of positive predictions.

3. Recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

- Measures the ability of the model to find all relevant instances (violent frames).

4. Fl-Score:

$$\text{F1-Score} - 2 \times \frac{\text{Preclston} \times \text{Recall}}{\text{Preclsion} + \text{Recall}} \quad (4)$$

- The harmonic mean of precision and recall, providing a single score that balances both metrics.

5. Structural Similarity Index (SSIM):

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_u + C_1)(2\sigma_{xp} + C_2)}{(\mu_x^2 + \mu_0^2 + C_1)(\sigma_2^2 + \sigma_3^2 + C_2)} \quad (5)$$

- μ_z and μ_y are the average of x and y respectively.
- σ_x^2 and σ_0^2 are the variances of x and y.
- σ_{xy} k the covariance of z and y.
- C_1 and C_2 are constants to stabilize the division.

6. Poak Signal-to-Noise Ratio (PSNF):

$$\text{PSNR} - 10 \times \log_{10}\left(\frac{(255^2)}{\text{MSE}}\right) \quad (6)$$

- Where MSE (Mean Squared Error) is calculated as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (I(i) - K(i))^2 \quad (7)$$

- $I(i)$ and $K(i)$ are the pixel values of the original and predicted frames, respectively, and N is the total number of pixels.

Alongside these classification measures, the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) are used to assess the validity of frame predictions. SSIM evaluates the perceived quality of anticipated frames in relation to actual frames by analyzing structural information, brightness, and contrast. A high SSIM score indicates that the predicted frames exhibit considerable visual resemblance to the real frames, which is essential for situations where visual accuracy is paramount. Likewise, PSNR quantifies the disparity between expected and actual pixel values; higher PSNR values indicate superior quality and more similarity across frames.

A. Implementation and Real-Time Testing

The method has a final application on a live video streaming system at the successful training and testing of the model. Indeed, this method can directly analyze video records totally; therefore, such violent events will be recognized with speed. The method is developed based upon the requirement to support a kind of continuous monitoring of video material with immediate notification for any detected violent elements. This capability makes initiatives regarding the protection of children from harmful material possible because it makes preventive action possible within a short time. Users are involved at this level of implementation. Collaboration with parents, educators, and child psychologists ensures the outputs from the model meet the expectations of practice and safety requirements. The stakeholders provide critical insights into the practical ramifications of the model's predictions, allowing modifications to the detection criteria and thresholds. They may emphasize certain behavioral or visual indicators that should be emphasized in the detection algorithm, ensuring the system is attuned to the subtleties of violent material.

The implementation phase also includes thorough testing to evaluate the model's performance across various real-

world circumstances. This entails assessing the model's reactions to several video categories, including varied genres, styles, and settings. Ongoing surveillance throughout this period is crucial to detect any possible deficiencies or opportunities for improvement. Should the model have difficulties with certain categories of violent material or produce an excessive number of false positives, modifications may be implemented to enhance its detection proficiency.

Furthermore, continuous data collecting during real-time testing facilitates the continued enhancement of the model. Expanding and diversifying the dataset via the collection of more samples of violent and non-violent material is essential for retraining and improving the model's accuracy and resilience. This iterative method cultivates a feedback and refinement loop, guaranteeing the system's efficacy in responding to the ever-changing multimedia content landscape.

The assessment and implementation stages are essential for the success of the video frame prediction and violence detection system. By using a complete array of measures to evaluate performance and soliciting stakeholder input, the model may be seamlessly incorporated into a real-time environment that emphasizes child safety. This dedication to ongoing enhancement and practical applicability establishes the system as an essential instrument in mitigating children's exposure to violent material in a more digital era.

TABLE II. RESULTS OF PROPOSED MODEL

Metric for Proposed Model ConLSTM	***Value***
Accuracy	92.5%
Precision	89.7%
Recall	91.2%
F1-Score	90.4%
SSIM	0.85
PSNR (dB)	32.1

Fig. 2. Performance Metrics of the Convolutional LSTM Model

Figure 2 graph illustrates the performance metrics of the Convolutional LSTM model throughout the training process, highlighting accuracy, precision, recall, and F1-score. As the number of epochs increases, all metrics demonstrate a consistent upward trend, indicating that the model is effectively learning to identify violent content in videos. Notably, the F1-score, which balances precision and recall, stabilizes at a high value, reflecting the model's ability to maintain a low rate of false positives and false negatives. This improvement across metrics underscores the model's robustness and its potential applicability in real-time video monitoring for child safety.

V. CONCLUSION

The proposed paper introduces a novel design of a new type of Convolutional LSTM architecture specifically targeting identification of automated content, especially related to violent and explicit material. As most proposals dealing with such urgency, the proposed approach particularly targets internet content moderation in relation to the protection of children. Now with this new multimedia stream content on YouTube and Facebook, there is a rising need for great strategies that keep the youth away from harming materials. The approach now includes deep learning techniques in addition to incorporating spatial and temporal data, which greatly boost the model's prediction and categorization of instances of violence.

From the results shown above, the model achieved very high precision, recall, and high F1-score at 92.5% accuracy, indicating that the model had successfully performed in this regard. Results suggest that it recognizes violent material correctly, thereby effectively reducing false positives and negatives that are crucial for its applications in child protection. Also, the use of quality metrics such as SSIM and PSNR ensures visually coherent predictions of the model, so this model has more significance in applications where responsiveness should be real-time. Deployment of the

trained model in a live video streaming environment allows effective and efficient detection of violent events and, hence, presents a useful solution to the parents, educators, and the content moderators. We could involve users at the deployment phase to better specify the criteria for detection and determine whether it meets expectation in reality, which are very important to us in developing a system that adjusts with the online changed nature of content. Continuous feedback and improvement are the keys.

For future work, more complex architectures, like transformers, can enhance the model and capture even more complex temporal patterns in video data. The multimodality may also be integrated, by incorporating audio and text to increase the accuracy of detection for content that is nuanced but would elude the visuals alone. There would be greater accessibility through parental controls and having real-time capabilities on edge devices, such as mobile phones. Standardized content moderation guidelines established in collaboration with social media platforms and policymakers could further enhance the safer digital space for these younger users.

REFERENCES

[1] D. Danier, F. Zhang, and D. R. Bull, "BVI-VFI: A Video Quality Database for Video Frame Interpolation," IEEE Transactions on Image Processing, vol. 32, pp. 6004–6019, 2023, doi: 10.1109/tip.2023.3327912.

[2] K. Suzuki and M. Ikehara, "Residual Learning of Video Frame Interpolation Using Convolutional LSTM," IEEE Access, vol. 8, pp. 134185–134193, 2020, doi: 10.1109/access.2020.3010846.

[3] "An Effective Video Event Classification by Optimizing the Hyper-Parameters Using Improved Pelican Optimization and Bi-LSTM Classifier," International Journal of Intelligent Engineering and Systems, vol. 17, no. 1, pp. 20–30, Feb. 2024, doi: 10.22266/ijies2024.0229.03

[4] R. Tyagi, GS Tomar, "Unfamiliar Sides, Video, Image Enhancement in Face Recognition", International Journal of Hybrid Information Technology, Vol. 9, No.11, pp.255-266, 2016.

[5] Amit Pandit, Shekhar Verma& G.S. Tomar, "Pruned AZB for reduced complexity block matching in video compression", IEEE International Conference Computer Information Systems and Industrial Management Applications (CISIM), 2010, pp 553-556, 2010.

[6] R. Gayathri, H. B. K, D. Sharmiladevi, Sajitha. L. P, V. S. Rao, and G. Nirmala, "Tracing the Motion in High-Speed Video Frames Using Deep Learning," 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), pp. 1–5, Nov. 2023, doi: 10.1109/rmkmate59243.2023.10368750.