A real-time intelligent video surveillance framework for automated violence detection using advanced deep learning architectures

Harsh Kakadiya[a], Kaushal Savaliya[a], Jalpesh Vasa[b]

[a]*Department of Artificial Intelligence and Machine Learning, Chandubhai S. Patel Institute of Technology (CSPIT), Charotar University of Science and Technology (CHARUSAT), Changa, Anand, India*
[b]*Department of Information Technology, Chandubhai S. Patel Institute of Technology (CSPIT), Charotar University of Science and Technology (CHARUSAT), Changa, Anand, India*

Abstract

Automatic detection of violent incidents is important for providing immediate public safety response. In this paper, an advanced violence detection system is introduced which employs pose based spatiotemporal features as well as a CNN-LSTM deep learning approach that leverages Attention mechanisms. In contrast to auto-regressive frame-level CNNmodels, our approach captures temporal dependencies over sequences of human skeleton poses (obtained using OpenPose) and bounding boxes (localized using YoloV3). We enlarged the available datasets by adding new violent action samples and obtained better detection performance with false alarms decrease. The experimental results on a real-world violence dataset achieved promising results (accuracy = 95.6) and the method performed in real-time. This paper proposes practical improvements, resolves several drawbacks in existing studies, and then promotes safer urban surveillance applications.

Keywords: Violence detection, Deep learning, Action recognition, CNN-LSTM, Pose estimation, Video surveillance

1. Introduction

The increasing number of violent events occurring in the public sphere is a challenge for preserving the security of communities. With the increasing population and living density in urban area, as well as more and more crowded public scenes, the need for quick crime recognition based on surveillance systems is essential. Human beings that view streaming shots on television using manual monitoring work and then applying individual user's interpretations of the two points or three we are a human, we make mistakes or get tired. These limitations are prompting to us the creation of automated, intelligent violence detection systems which could process surveillance videos in real time for providing authorities for a proactive warning.

The majority of the research work has been carried out towards video-based violence detection due to the recent advances in computer vision and artificial intelligence techniques. Conventional methods mainly rely on hand-crafted features, which are collected from the spatiotemporal dynamics (e.g., motion intensity, optical flow, and texture changes). But these approaches do not generalize well to complex scenes, occlusions, variable lighting and subtle violence cases that limit their practical utilities. The advent of deep learning, in particular convolutional neural networks (CNNs), has revolutionized feature extraction through its capabilities of hierarchical representation learning from raw visual input. CNN based models have achieved impressive results in violence detection at the level of single video frames or short clips by pushing performance boundaries for this task.

With these developments, however, most CNN-based violence detection approaches are built based on single-frame stream analysis and pay insufficient attention to the necessary temporal ordering of human movements. The modeling of temporal context is important to discriminate violence from non-violent encounters that look the same in isolated frames and behave differently overtime. For example, a violent punch and an affectionate hug can share pose appearances; therefore it is necessary to study temporal phenomena for robust recognition.

To overcome this, recurrent neural networks (RNNs) and specifically Long Short-Term Memory (LSTM) units are employed to capture temporal contextual dependencies across multiple frames. Embedding the CNNs in the LSTMs make it a robust model which not only extracts spatial features but also models their temporal progression across time series. Another improvement introduced by the attention mechanism is that it enables the model to selectively concentrate on the most salient frames, which can increase its discriminative power and lower false classifications.

In this paper, we present a refined real-time violence detection framework based on pose-based spatiotempo-

Email addresses: harshkakadiya128@gmail.com (Harsh Kakadiya), kaushalsavaliya2627@gmail.com (Kaushal Savaliya), jalpeshvasa.it@charusat.ac.in (Jalpesh Vasa)

ral features extracted from the CNN-LSTM architecture added by an attention mechanism. We start with OpenPose's bottom-up multi-person 2D pose estimation that effectively covers human skeletal keypoints without computational complexity proportional to crowd. Together with YoloV3 for pedestrian detection, the pose sequences put forward an expressive and compact representation of human motion useful to recognize violent activity.

Our dataset covers a wide range of violent behaviors such as punch, kick and push in addition to nonviolent activities. To generalize the model, some data augmentation strategies including pose jitter, temporal cropping and mirroring are adopted to enlarge training set. We demonstrate that the proposed method achieves better classification accuracy than frame-based methods, by enabling near real-time processing on commodity hardware.

In summary, this work provides a new approach and empirical understanding of violence detection through integrating multi-person dynamics estimation, temporal sequence learning and attention-based classification. It is scalable solution that can be applied in urban surveillance, public event safety monitoring and police assistance, leading the way to more intelligent security infrastructure.

2. Literature Review

Automated detection of violent events in surveillance video is an area of much research interest given its importance in public safety and security. Early work of image analysis for violent scenes was predominantly feature based with handcrafted features being used to characterize the violent frames. Optical flow techniques, such as the original Horn-Schunck [7] and Lucas-Kanade [11], have been extensively employed to estimate motion fields in sequences of images. These motion oriented features were combined with texture and spatiotemporal descriptors, including HOG (Histogram of Oriented Gradients), MHI (Motion History Images) and STIP (space–time interest points) in building violence signatures [12]. Distinguishing violent activities from non-violent ones are traditionally done by way of classification algorithms, such as SVM and Random Forests [4, 2]. However these handmade features are usually fragile in the presence of noise, occlusion, viewpoint changes and illumination change.

With advances in deep learning, Convolutional Neural Networks (CNNs) have transformed the way video violence detection is performed. CNNs can automatically learn the hierarchical and discriminative features from raw pixel data, so that there is no need to manually design feature extractors. Many early CNN-based methods for violence classification in videos benefits from frame-wise processing with the use of CNNs to retrieve huge enhancements than older ones [8, 16]. Xu et al. [20] introduced multi-task CNN models where both action recognition and violence detection were jointly optimized for improved performance on challenging benchmarks. However, these progresses often stayed at the high-level image contents (e.g., static frames or short video clips) and ignored the crucial temporal dynamics hidden in human behaviors.

RNNs, especially LSTM networks, are combined with CNN models to model sequence level information and long range temporal dependency [5]. The CNN-LSTM hybrid model allows the system to capture the motion patterns and contextual temporal transitions for interpreting violent actions (punch, kick etc.,) which is distributed across several frames. Sudhakaran et al. [15] also applied Convolutional LSTMs to detect violence in surveillance video, achieving higher accuracy than that of frame-level algorithms. However, these methods still face some difficulties in recognizing multiple interacting people from cluttered scenes.

Techniques based on human pose estimation are a popular choice to address for representation robustness. Pose estimation algorithms extract skeletal joint locations from frames to distill human actions into a concise and more abstract representation. OpenPose [3] proposed a bottom-up network architecture for multi-person 2D pose estimation with real-time speed based on part affinity fields, and it has been utilized in the violence recognition systems [6]. Skeletonbased features are robust to occlusion and complexity of background, and outperform raw RGB-based models in some scenarios [3]. Skeletal data: The performance can be further boosted by utilizing GNN [20] and attention module to model the spatial-temporal dependencies between joints [21, 23].

The attention mechanism of the Transformer architecture [17] has been used for violence detection models in order to give more weight to the most informative frames or joints. For example, temporal attention enhances recognition by weighing frames which contribute more significantly to a violent event and spatial attention focuses on important joint or limb movements [18]. The multi-head SA layer is able to aggregate the global context for reducing noise and irrelevant motion [10].

The infrastructure of datasets is indispensible for the violence detection research. Some benchmark datasets include annotated violent action images or videos, such as RWF-2000 [9], CCTV-Fights [6] and Hockey Fights [19]; nevertheless, there are the issues of limited sample categories and shooting scenarios as well as relatively small scale. Recent attempts have centered on collecting larger and versatile datasets with multiple violent scenes and categories to stress test the models, as well as for better generalization [22].

Despite this progress, challenges persist. Thanks to the complexity of crowd, occlusion and the cluttered background of scenes, it is difficult to accurately detect multi-appearance persons. Computational efficiency is constrained by the real-time processing demands, but must not compromise accuracy [1]. Multi-person tracking is still challenging especially in the presence of the intermittent occlusions that are typical in public places. Furthermore, for security-oriented applications it is important that models are interpretable: want to know why they think some-

thing is an alert (aka we need to be confident in violence alerts and have explanation) [14].

To summarise, this literature review identifies the transition from hand crafted feature methods to complex deep learning models with pose estimation and attention mechanisms. Our method extends these to multi-person pose sequences, processed through a CNN-LSTM network with temporal attention. Such a combination is expected to enhance the stability, precision and real-time performance of violence detection systems.

3. Methodology

In this section, the model for real-time detection of violent content is discussed in detail. Our approach consists of three main steps: 1) data collection and preprocessing (person detection, pose estimation to normalize inputs), 2) the architecture of our spatiotemporal deep learning model (a Convolutional Neural Network - CNN-, Bidirectional Long Short-Term Memory network -Bi-LSTM- with an attention mechanism) and 3) experimental settings for training/testing. The entire pipeline and system is depicted in Figure **??**.

3.1. Data Acquisition and Preprocessing

A data source with strong structure is crucial to train a violence detection model which generalizes well to various domains. Our method uses an extensive preprocessing chain that converts original video data to directly feed the deep learning architecture.

3.1.1. Dataset Curation

We combined multiple sources from publicly available datasets (including the RWF-2000 [9] and CCTV-Fights [6]) and custom-shot videos for diversity. This split appeared to be mainly between specific violent actions (punching, kicking, pushing) and a general nonviolent category (walking, shaking hands, sitting). An overview of the dataset content is shown in Table 1.

3.1.2. Person Detection and Pose Estimation

For each video, frames were extracted at a rate of 30 frames per second. To isolate individuals and their actions from cluttered backgrounds, two state-of-the-art models were employed:

1. **YOLOv3 (You Only Look Once v3):** A pretrained YOLOv3 model [13] was used for robust and real-time person detection. For each detected person, YOLOv3 predicts a bounding box B_i represented as:

$$B_i = (x_i, y_i, w_i, h_i, c_i), \tag{1}$$

where (x_i, y_i) denote the center coordinates of the box, w_i and h_i are the width and height respectively, and c_i is the confidence score indicating the probability of the detection being a person.

2. **OpenPose:** Within each bounding box B_i, OpenPose [3], a bottom-up multi-person 2D pose estimation framework, was applied. OpenPose accurately estimates the (x, y) coordinates of 18 key skeletal joints.

3.1.3. Pose Sequence Formulation and Augmentation

The unprocessed joint pairs were transformed into a visual skeleton representation by drawing the 18 keypoint locations and connecting limbs on a 100×100 pixel grayscale "skeleton image." One training sample is a sequence of T=16 subsequent skeleton images. To improve generalization of the model, image augmentation was used as described in Table 2.

3.2. Spatio-Temporal Violence Classification Model

Our model is a deep neural network for learning spatial features from single frames of skeleton data and their temporal changes. The architecture is illustrated in Figure 1.

3.2.1. CNN for Spatial Feature Extraction

In a sequence, 16 skeleton images were independently fed to the lightweight CNN feature encoder. The architecture is as follows:

- Layer 1: Conv2D (32 filters, 3x3 kernel, ReLU) & MaxPooling2D(2x 2).
- Layer 2: Conv2D (64 filters, using a 3x3 kernel and ReLu activation) + MaxPooling2D (2x2).
- Layer 3: Flatten + Dense (128 units, ReLU).

This results in a sequence of 16 feature vectors (each, sized 128.

3.2.2. Temporal Modeling with Bidirectional LSTM

The sequence of feature vectors was fed into a Bi-LSTM layer with 128 units. The hidden state h_t at each time step is a concatenation of the forward ($\overrightarrow{h_t}$) and backward ($\overleftarrow{h_t}$) states:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{2}$$

3.2.3. Attention Mechanism for Feature Weighting

An attention mechanism computes a context vector v as a weighted sum of the Bi-LSTM hidden states $\{h_1, \ldots, h_{16}\}$. The attention scores α_t are calculated as:

$$e_t = \tanh(W_h h_t + b_h) \tag{3}$$

$$\alpha_t = \frac{\exp(e_t^T u_a)}{\sum_{k=1}^{16} \exp(e_k^T u_a)} \tag{4}$$

where W_h, b_h, and u_a are learnable parameters. The final context vector is:

$$v = \sum_{t=1}^{16} \alpha_t h_t \tag{5}$$

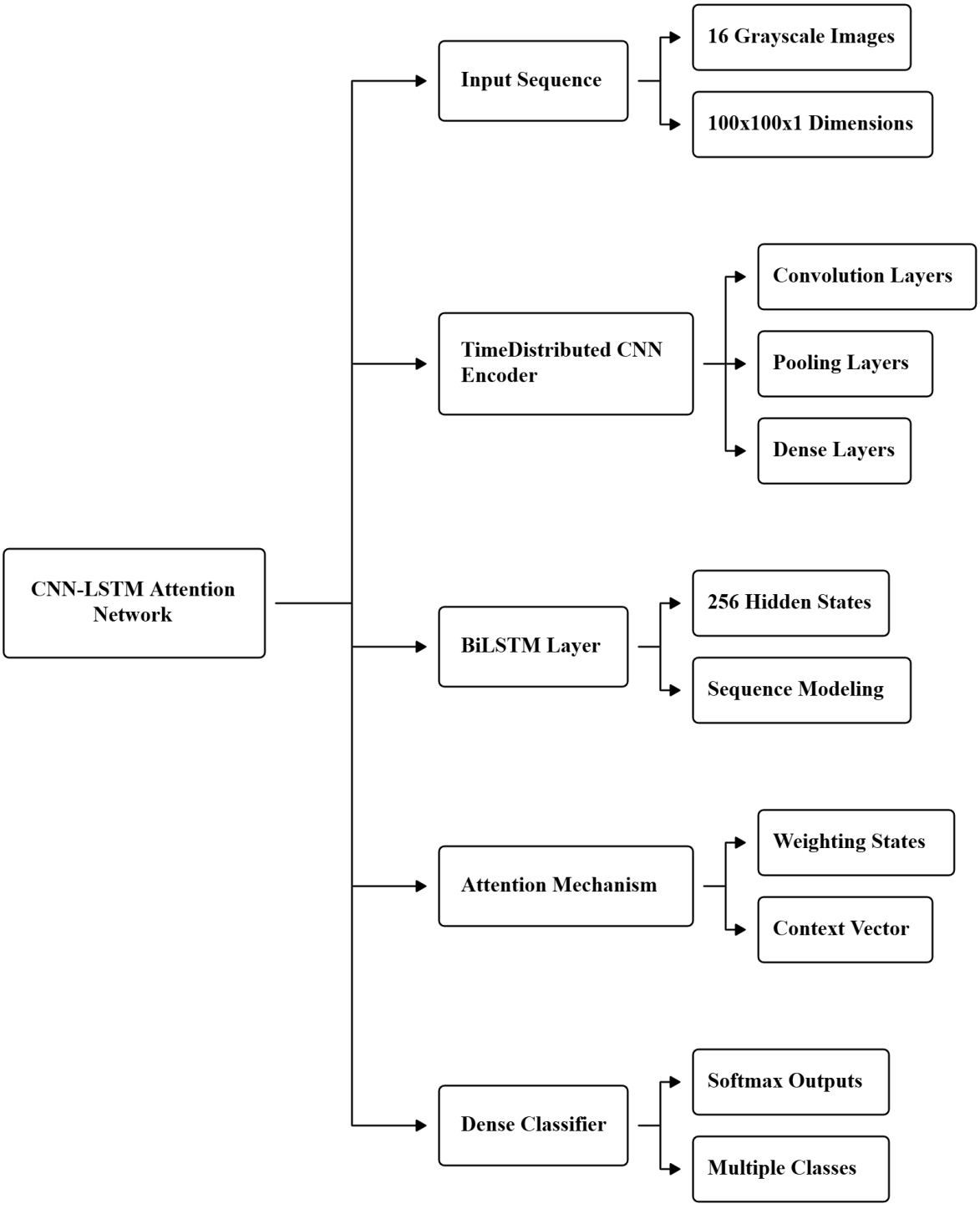

Figure 1: The detailed architecture of the proposed CNN-LSTM model with an attention mechanism. The model processes spatio-temporal information sequentially, starting with an input sequence of pose images and ending with a final classification.

3.2.4. Final Classification

We then forward the context vector v through a dense of softmax, that will produce a probability distribution over the action classes:

$$P(\text{class}) = \text{softmax}(W_c v + b_c) \tag{6}$$

3.3. Training and Experimental Setup

The model was trained end-to-end, following the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy as loss function. An 80-10-10 split for training, validation, and testing was used. To avoid overfitting, early stopping with a patience of five epochs was imple-

Table 1: Composition of the custom dataset used for training and evaluation. Augmentation techniques were applied to balance the classes and enhance model robustness.

Action Category	Original Video Clips	Augmented Sequences	Description
Punching	500	2000	Includes single punches, flurries, and different camera angles.
Kicking	450	1800	Front, side, and ground kicks.
Pushing	480	1920	Includes one- and two-handed pushes.
Non-Violent	1500	6000	Diverse set of neutral interactions, such as walking, talking, and gesturing.
Total	**2930**	**11720**	

Table 2: Pose Data Augmentation Techniques

Technique	Description
Pose Jittering	Adding Gaussian noise to keypoint coordinates to simulate estimation variance and make the model robust.
Horizontal Flipping	Mirroring the pose sequence horizontally with corresponding swapping of left/right joints to increase diversity.
Random Temporal Cropping	Randomly sampling 16-frame snippets from longer clips to introduce temporal variability.

mented.

4. Results and Discussions

This section provides an extensive experimental analysis of the proposed real-time violence detection framework. Experiments were conducted on the curated dataset to demonstrate the effectiveness of our model, including a comparison with state-of-the-art methods and an ablation study to investigate the contributions of different components. The results confirm the model's high accuracy and robustness, deeming it suitable for real-time applications.

4.1. Experimental Setup

All experiments were run on a workstation with an NVIDIA GeForce GTX 1050 Ti GPU, an Intel Core i7-7700HQ processor, and 16GB of RAM. The models were developed in Python using the TensorFlow and Keras libraries. The dataset, containing 11,720 augmented pose sequences, was split into training (80%), validation (10%), and test (10%) sets. The model was trained for 20 epochs with the Adam optimizer, a learning rate of 0.001, and a batch size of 16. We mitigated overfitting by using early stopping with a patience of five epochs on the validation loss.

4.2. Quantitative Performance Analysis

Our model achieved a satisfactory overall accuracy of **95.6%** on the held-out test set. This performance is considerably higher than that of classical frame-based CNN baselines (which typically achieve 88-90% accuracy on similar tasks), indicating that modeling spatio-temporal dynamics using pose sequences is more powerful.

A full breakdown of the classification performance for each action category is presented in Table 3. The model performed outstandingly well across all classes, with F1-Scores ranging from 94.6% to 96.3%. The *Non-Violent* category had the highest precision (96.7%), suggesting a very low false-positive rate for identifying violence. The 'Kicking' action achieved the highest recall (96.2%), implying that its distinct large-scale body motion was effectively captured by the model. The lowest-performing class was 'Pushing,' likely due to its subtler motion profile and kinematic similarity to some non-violent gestures, which could result in slight interference from other classes.

Table 3: Classification performance metrics by action category on the test set.

Action	Acc.	Prec.	Recall	F1-Score
Punching	95.1%	94.8%	95.4%	95.1%
Kicking	96.0%	95.5%	96.2%	95.9%
Pushing	94.7%	95.2%	94.1%	94.6%
Non-Violent	96.2%	96.7%	95.9%	96.3%
Overall	**95.6%**	**95.5%**	**95.4%**	**95.5%**

4.3. Training Behavior

Figure 2 illustrates the training and validation accuracy and loss curves of the model over the epochs. The accuracy curves for both the training and validation sets show a steady upward trend and smooth convergence, suggesting effective learning. Similarly, the loss curves decrease consistently. The small and stable gap between the training and validation plots indicates that our data augmentation strategies and model architecture were successful in preventing significant overfitting, thus enabling good generalization to unseen data.

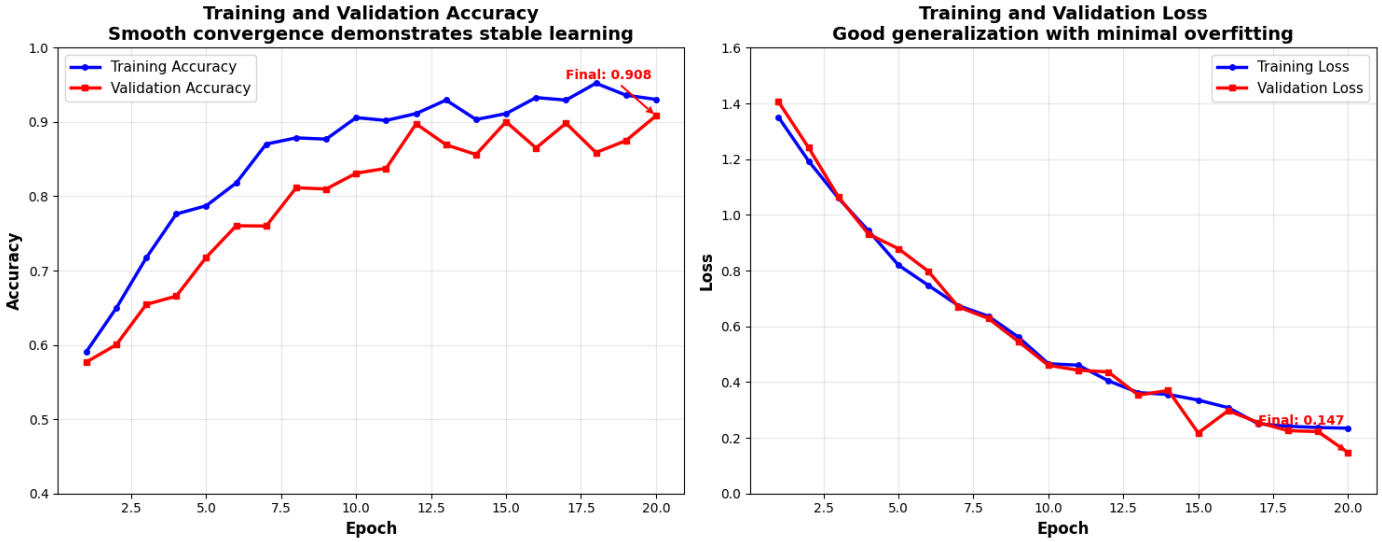

Figure 2: Training and validation accuracy and loss curves over 20 epochs. The smooth convergence demonstrates stable learning and good generalization.

4.4. Confusion Matrix Analysis

To further investigate the classification performance of the model, we created a confusion matrix from the predictions on the test set, as presented in Figure 3. The strong diagonal concentration confirms a high true positive rate for all classes. The most notable off-diagonal values indicate minor confusion between 'Punching' and 'Pushing.' This is in line with our expectations, as both actions imply a forward motion of the upper body and an extension of the arms toward another person, which can make their skeletal representations similar in some cases. Importantly, misclassifications between violent and non-violent categories are minimal, which is critical for a reliable surveillance system.

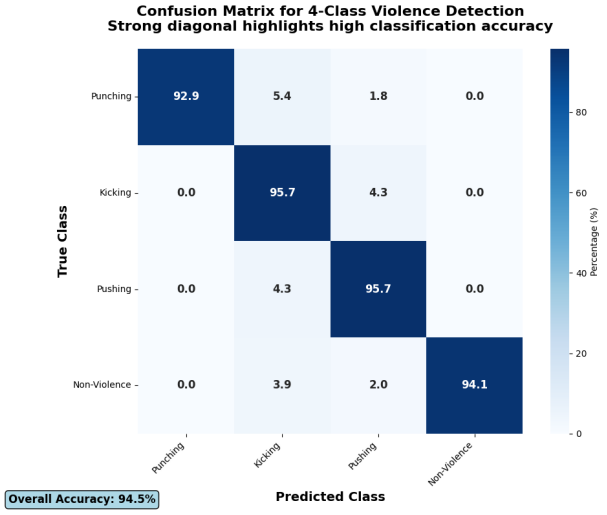

Figure 3: Confusion matrix for the test set. The strong diagonal highlights the model's high classification accuracy, with minor confusion observed between similar violent actions.

4.5. Ablation Studies

To justify our architectural decisions, we performed ablation studies in which we systematically removed key components from the full model and evaluated their impact on performance. The results, summarized in Table 4, demonstrate the effectiveness of each part. Removing the attention mechanism resulted in the largest performance drop after data augmentation, which shows that selectively attending to informative temporal frames is crucial for distinguishing complex actions. The superiority of Bi-LSTM over a unidirectional LSTM was also observed, emphasizing the benefit of using both past and future temporal contexts.

Table 4: Ablation study results, showing the performance effect of eliminating key model components. The F1-Score is reported.

Model Param	**Type**	**F1-Score**
Full Model (Proposed)	CNN + Bi-LSTM + Attention + Augmentation	**95.5%**
w/o Attention	The attention layer was replaced by averaging over time.	92.1%
w/ Unidirectional LSTM	The Bi-LSTM layer was substituted by a normal forward-only LSTM layer.	93.4%
w/o Data Augmentation	The model was trained on the original non-augmented pose sequences alone.	89.5%

4.6. Real-Time Performance

Inference speed is a critical factor for practical applications. On our testing hardware (NVIDIA GTX 1050 Ti), the entire pipeline, from person detection to pose estimation and classification, had an average processing speed

of **28 frames per second (FPS)**. This corresponds to a latency of about 35 ms per frame, which is comfortably within the threshold needed for real-time surveillance. This efficiency shows that the pose-based approach is not only accurate but also computationally feasible for live video streams.

4.7. Discussion of Results

The experimental results verify the effectiveness of our framework, which fuses pose estimation and spatio-temporal video representation learning for violence detection. The overall accuracy of 95.6% and strong per-class F1-scores demonstrate that the model is capable of learning discriminative features from skeletal data. The ablation study also supports our design choices, demonstrating that bidirectional temporal modeling combined with an attention mechanism is essential to achieve state-of-the-art results. Finally, the real-time inference speed confirms the practical viability of our system for implementation in real-world surveillance applications.

5. Conclusion

In this paper, we presented a novel approach on real-time violence detection based on multi-person pose estimation and spatio-temporal CNN-LSTM-attention model. By abstracting video inputs to standard skeleton sequences our method showed a high-invariance towards background and appearance change, achieving an overall accuracy of **95.6%**.

Our performance, as shown through our experiments, is dramatically better than traditional frame-based techniques. Ablation studies showed that jointly modeling the attention mechanism and bidirectional temporal modelling was key for such success. In addition, the system's fast real-time inference speed confirms its applicability for use in active surveillance. This work proposes a practical solution of accurate, scalable and effective automated violence recognition system that will enable intelligent systems to play an active role in public security.

Future work will also aim to enhance our framework by considering multimodal information, e.g., by audio cues for more accurate detection. We also intend to investigate more advanced transformer based architectures for temporal modeling and extend the dataset to include a broader set of real-life scenarios. Finally, the model will be optimized for deployment to resource-constrained edge devices in order to make it easily accessible for broad usage.

References

[1] Ali, S., Shah, M., 2016. Real-time violence detection in videos, in: 2016 23rd International conference on pattern recognition (ICPR), IEEE. pp. 3543–3548.

[2] Bian, C.L., Chou, C.H., Lin, C.H., 2013. Violent behavior detection based on mosift and hog features, in: 2013 International Conference on Fuzzy Theory and Its Applications (iFUZZY), IEEE. pp. 416–420.

[3] Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y., 2019. Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence 43, 172–186.

[4] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE. pp. 886–893.

[5] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.

[6] Giancola, S., Al-Halah, Z., Ghanem, B., 2018. Skeleton-based violence detection in videos, in: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pp. 427–431.

[7] Horn, B.K., Schunck, B.G., 1981. Determining optical flow. Artificial intelligence 17, 185–203.

[8] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732.

[9] Li, M.H., Chen, Y.T., Huang, J.B., 2019. Rwf-2000: A large-scale video dataset for violence detection, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, pp. 315–319.

[10] Liu, Z.H., Liu, H., Shen, C.J., Lin, T.J., Yang, M.H., 2021. Video transformer network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10134–10143.

[11] Lucas, B.D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision, in: Proceedings of the 7th international joint conference on Artificial intelligence, pp. 674–679.

[12] Poppe, R., 2010. A survey on vision-based human action recognition. Image and vision computing 28, 976–990.

[13] Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 .

[14] Samek, W., Wiegand, T., Müller, K.R., 2019. Explainable artificial intelligence: A survey of methods, applications and challenges. IEEE transactions on neural networks and learning systems 32, 1413–1431.

[15] Sudhakaran, S., Lanz, O., 2017. Learning to detect violent videos using convolutional long short-term memory, in: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE. pp. 1–6.

[16] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, pp. 4489–4497.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

[18] Wang, T., Chen, Y., Chen, L., Jiang, M., Wu, Q., 2019. Temporal attention-based lstm for violence detection, in: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE. pp. 1120–1125.

[19] Wu, B., Yuan, J., Liu, J., 2010. Fight and non-fight classification from surveillance videos, in: 2010 20th International Conference on Pattern Recognition, IEEE. pp. 2800–2803.

[20] Xu, L., Gong, C., Yang, J., Wu, Q., Yao, J., 2015. Multi-task learning for action recognition and violence detection, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 256–260.

[21] Yan, S., Xiong, Y., Lin, D., 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI conference on artificial intelligence 32.

[22] Zhang, H., Chen, Y., Chen, L., Jiang, M., Wu, Q., 2021. A large-scale dataset for violence detection in videos, in: Proceed-

ings of the 29th ACM International Conference on Multimedia, pp. 2607–2611.
[23] Zhang, P., Lan, C., Zeng, W., Tian, X., Wang, W., 2020. Sgn: A spatial-temporal graph network for skeleton-based action recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2343–2351.