Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020)
IEEE Xplore Part Number: CFP20P17-ART; ISBN: 978-1-7281-5821-1

Deep NeuralNet For Violence Detection Using Motion Features From Dynamic Images

Aayush Jain, Dinesh Kumar Vishwakarma
Department of Information Technology
Delhi Technological University
Delhi-110042, India
aayush89j@gmail.com, dvishwakarma@gmail.com

***Abstract*—Violent Action Recognition has been a challenging topic in monitoring human activities, especially in public premises. Deep neural nets and Transfer learning have proven highly successful in the detection of violent activities. In this paper, a novel deep NeuralNet system is proposed for the task of Detecting Violence by extraction of motion features from RGB Dynamic Images (DI). Motion feature extraction and prediction of violent content using a stream of RGB DI is done effectively by fine-tuning the pre-trained Inception-Resnet-V2 model. Improved and highly effective methods for intelligent analysis are highly demanded. For performance validation of the proposed novel model, tests are performed on three popular and publically available benchmarks – Hockey Fight dataset, Real Life Violence Dataset, and movie dataset.**

***Keywords*—Violence Detection; Deep NeuralNet; Transfer Learning; Dynamic Images; Inception-Resnet-V2; CNN;**

I. INTRODUCTION

With the spike in violent activities like the recent Delhi riots [1] or the recent mass protests in the US resulting in violent events [2], it is becoming more and more necessary to develop efficient and highly accurate automated intelligent systems which can analyze the public premise for detecting and avoiding suspicious events and violent activities for ensuring better public safety and enabling agencies in deploying better security measures. Recent state-of-the-art has been quite successful and effective in creating surveillance systems and detection methods which can enable enforcement agencies to avoid such incidents [3]–[6]. CNN's have gained remarkable success in recent years in the field of image and video classification with incomparable accuracy and modeling capacity [7]–[9]. Prior studies in the area of violence detection [10]–[16] have ascertained that temporal, spatial, and motion information act as an important trait to classify and detect violent activities in surveillance footage and real-life situations in an effective manner. Researchers have also done extensive work in detecting violent activities using a single-stream architecture extracting spatial information or temporal information or motion information and also by multi-stream architectures to extract spatial-temporal or spatial-temporal-motion data that has improved CNN based systems accuracy by examining RBG and depth based appearance and motion details.

Yet most of the approaches like [17] do not necessarily model the dynamics of the video. RGB images, extracted as a stack of frames from the video, are used to learn discriminative attributes that allow the model to obtain the highest accuracy in tasks of recognition and classification since the primary aim is to represent the action class rather than the motion. Whereas motion features contain a decent amount of evidence for distinguishing the movement and human action. In this regard, optical flow [14], [18]–[20], and dense trajectories [21] are used extensively for illustrating the movement of the entity in videos. Various methods proposed descriptors, based on optical flow orientation and magnitude changes for example ViF [22], [23], OViF [24], HOF and HOG [25] , HOMO [10], etc. However, these approaches regardless of being beneficial also have some disadvantages like high computational time limiting the possibility of real-time implementation of such systems. Similarly, dense trajectories can be extremely sensitive to camera viewpoint.
Considering all such approaches and their shortcomings [26] proposed a new approach which is an extremely powerful, simple, and yet efficient depiction of the video called Dynamic Images, summarizing the motion dynamics present in the video into a single image. Dynamic Images (DIs) can encrypt information in a conventional content-agonistic manner which results in highly effective long-term, stable representation of motion for classification and recognition tasks and also for other tasks like human pose estimation [27], [28], as well. Key benefits of using Dynamic Images which are as follows:

1. DIs can be attributed to various forms of sequences.
2. DIs are very efficient, fast, and simple.
3. Extraction of DIs reduces computational time, also reduces video analysis to single RBG image analysis.
4. Compact Representation of video.
5. DIs can be useful for large scale indexing
6. DIs can be processed by any CNNs

The sole objective of this paper is to emphasize the distinct

motion features embedded and encoded as RGB-DIs and how a deep learning neural net can be used to capture such motion characteristics and improve the task of violence detection. The fine-tuned pre-trained Inception-Resnet-V2 model is used to evaluate how well the proposed framework performs and establish the novelty of the work on the three publically available datasets or benchmarks – hockey fight, real-life violence, and movie dataset.

The paper is structured as follows: Description of the relevant work is discussed in section II, emphasizing the contribution in the field of violence detection. Section III deals in detail with the proposed framework and the results of conducted experiments are present in section IV. Lastly, concluded in Section V.

II. Related Work

Violence Detection systems dependent on deep learning like [29]–[31] are burgeoning with ever-increasing accuracies and advantages over the other. The evolutions of single-stream architectures into multi-stream state-of-the-art have also shown a significant increase in performance since these architectures incorporate the amalgamation of different information cues – motion, spatial, and temporal. Previous works based on extracting spatiotemporal features from video or pyramid of RGB image frames extracted from videos have shown tremendous potential in detecting violence with accuracies nearing almost 100% mark. [32] extracted frame-level characteristics from video using AlexNet as CNN followed by LSTM that uses convolutional gates, aggregating the frame-level characteristics and [33] used pre-trained light-weight MobileNet to reduce bulk processing of useless frames followed by 3D CNN for feature extraction. [6] used spatiotemporal encoder and bidirectional ConvLSTM architecture for better performance on complex and heterogeneous datasets like Violent Flow datasets, [34] also used CNN and LSTM architecture for extracting spatial features and learning temporal relation respectively resulting in outperforming speed as compared to previous models. These advancements are majorly based on spatiotemporal feature extraction in reaching extremely high accuracy. But with increasing complexities in such multi-stream architectures, it is also necessary to improve the quality of information that each stream captures.

Motion information is a rich tool for gathering knowledge dependent on human action. Techniques such as optical flow, optical flow based ViF, OViF are used to capture motion information. Optical flow-based methodologies record the optical-flow within consecutive image frames encapsulating the motion using key components. [22] used ViF descriptors, object detection method, for representing stats collected for short frame sequences, and bags of feature method for feature extraction. The fundamental aim of the strategy is to identify the transition of violent to non-violent action having the shortest delay since the change occurred. [24] used OViF for feature extraction and captures the essence of information about the motion magnitude change. However, only the local dynamics are provided by these methodologies, and local motion analysis is performed using simple summarization techniques. CNN can extract meaningful information from the input provided hence it is highly important to decide how the video information is provided to CNN. Recent advancements in capturing the motion information using Dynamic Images for the task of action recognition [35] have shown significant improvement in the action recognition task using motion information. Dynamic Images can summarize the whole video content into a single image (fig.1).

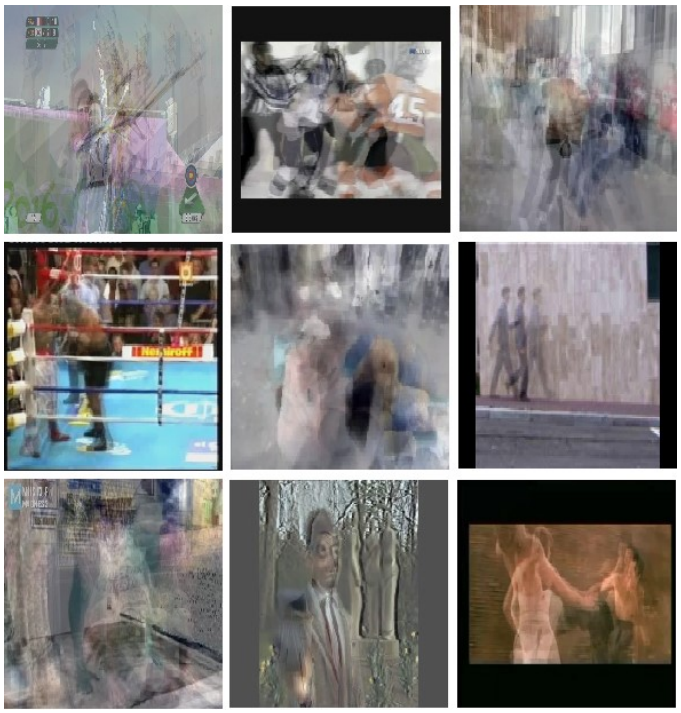

Fig. 1: Dynamic Images captured from 3 datasets - Hockey Fight dataset, Real-Life Violence Dataset and Movie Dataset summarizing violent and non-violent behavior

III. Proposed Methodology

The proposed deep neural net architecture, fine-tuned pre-trained Inception Resnet V2 based CNN model for Violence Detection is shown in Fig.2. The architecture is designed to learn motion features and use motion information for the task of Violence Detection. The motion content or the motion information present in the video is captured by transforming RGB video into Dynamic Image (DI). The transfer learning approach is used for recognizing violent and non-violent acts with the assistance of fine-tuned pre-trained Inception-Resnet-V2. DIs typically focuses on the salient object's movement by combining and averaging the background pixels as well as the movement patterns while retaining the long-term kinetics. The shape of the obtained DI is identical to the original frame. The constructed DI is transferred through the implemented architecture to compute the motion characteristics.

The architecture is a combination of Inception ResnetV2, pre-trained on the ImageNet Dataset, followed by a set of Dense and Dropout Layers. ReLU is used as the activation function on the Dense Layers expects the end Dense Layer. Last Dense Layer used the Softmax Activation. The

architecture of our proposed model is as follows :

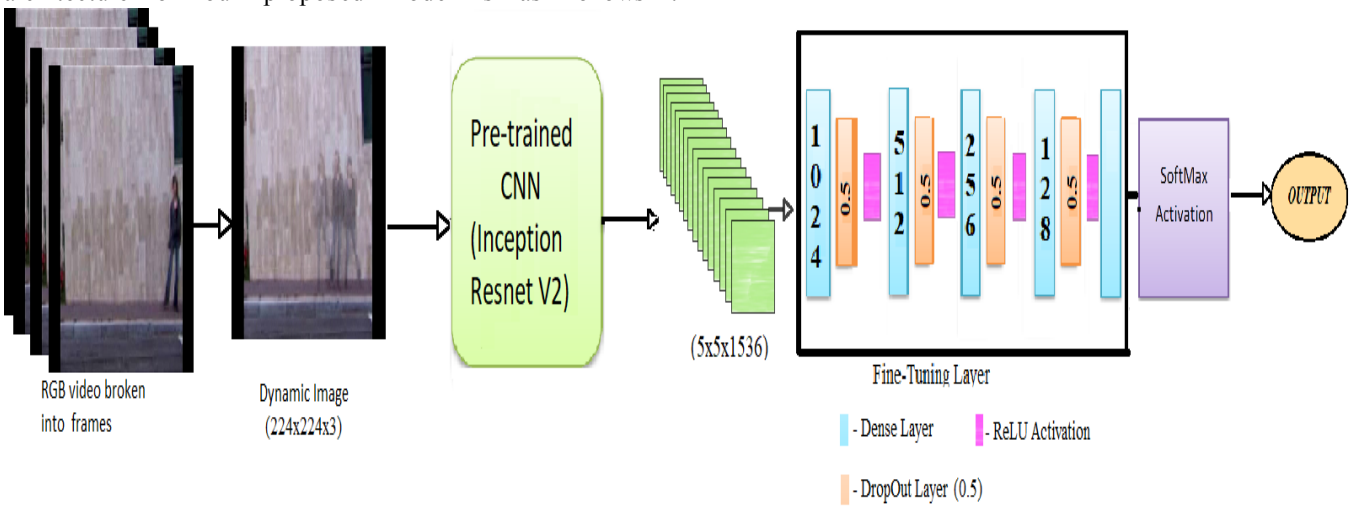

Fig. 2: Schematic Block Diagram of Proposed Framework

Input DI (224x224x3) – Inception-Resnet-V2 () – Intermediate Vector Shape (5x5x1536) – Dense Layer (1024) – DropOut Layer (0.5) – ReLU – Dense Layer (512) – DropOut Layer (0.5) – ReLU – Dense Layer (256) – DropOut Layer (0.5) – ReLU – Dense Layer (128) – DropOut Layer (0.5) – ReLU – Dense Layer (2) – Softmax ()

The pre-trained Inception-Resnet-V2 layers are not trained and kept frozen during the whole training duration. The layers for fine-tuning the model are trained end-to-end for updating weights according to the training sample. Since the data size is small, overfitting will be an issue. To counter it, the Dropout Layer helps to handle the overfitting phenomena as well as the training data size was also doubled by capturing the 2 dynamic images from a single video, created using distinct mutually exclusive frames meaning the frame used to create first DI was not used for creating the second DI. Early stopping was used so that model doesn't over-train. Without early stopping the validation loss increases and accuracy degrades. Model Checkpoint is used to obtain the best-trained model weights during training having the lowest validation loss. For testing purposes, the same saved model checkpoint having the best weights is used to achieve the desired results.

IV. Experimental Results

For substantiating the achievement of our proposed violence detection framework, three publically available benchmarks – Hockey Fight Dataset, Real Life Violence Dataset, and Movie Datasets are used.

A. Hockey Fight Dataset

The Hockey Fight Dataset was introduced by [36] containing 1000 violent and non-violent action clips each, recorded during a hockey game of the NHL. Each video has a frame rate of 25fps consisting of 50frames of 720x576 pixels. The dataset was labeled into a fight and non-fight category. All the clips have the same background with only Ice Hockey players entering the frames.

B. Real-Life Violence Dataset

A new and quite challenging benchmark is introduced [37], The Real-Life Violence Dataset, as compared to Hockey Fight and movie dataset with 2000 videos divided into 1000 violent and non-violent genre videos, having a wide range of gender, age, and race collected from diverse backgrounds and environment. Violent videos are captured in an environment like a prison, street, and schools whereas the non-violent videos are captured in environments like sports fields and arena featuring events such as swimming, doing archery, playing basketball. The average duration of the clip is 5seconds. With maximum duration being 7seconds and a minimum duration of 3 seconds with a frame rate of 25fps.

C. Movie Dataset

The Movie Dataset [36] is a considerably small dataset containing only 200 movie clips from action scenes bifurcated equally into violent and non-violent video genres. The movie dataset has also varying backgrounds. The clips have an average duration of 1 second and some clips have a duration of 2seconds, having a frame rate of either 25fps or 30fps.

In the experiment, the end-to-end training of fine-tuned layers of the proposed model takes place. Before the training phase begins, the dataset is subdivided into testing and training samples using the 30-70 splitting strategy, and training sample videos are processed to get dynamic images. For each training video, 2 dynamic images are computed by adopting the methodology proposed by [35]. RGB video was broken into a stack of frames and 2 sets of frames were produced, one starting with 1st frame having stride of frameRate/2 and the other set starting with 1st frame = frameRate/4 and having stride of frameRate/2. Each frame in the 2 sets was taken and split into R, G, and B channels separately. A coefficient was computed for each frame and a list of frames split by channels was multiplied by the coefficients. The weighted aggregate of all 3 channels - R, G, and B in each frame separately is collected to obtain the R, G, B channel of the DI. Once the R, G, B channels of DI is achieved, they were then merged to form a single DI

normalized into pixels having a value between 0-255 as shown in Fig. 3.

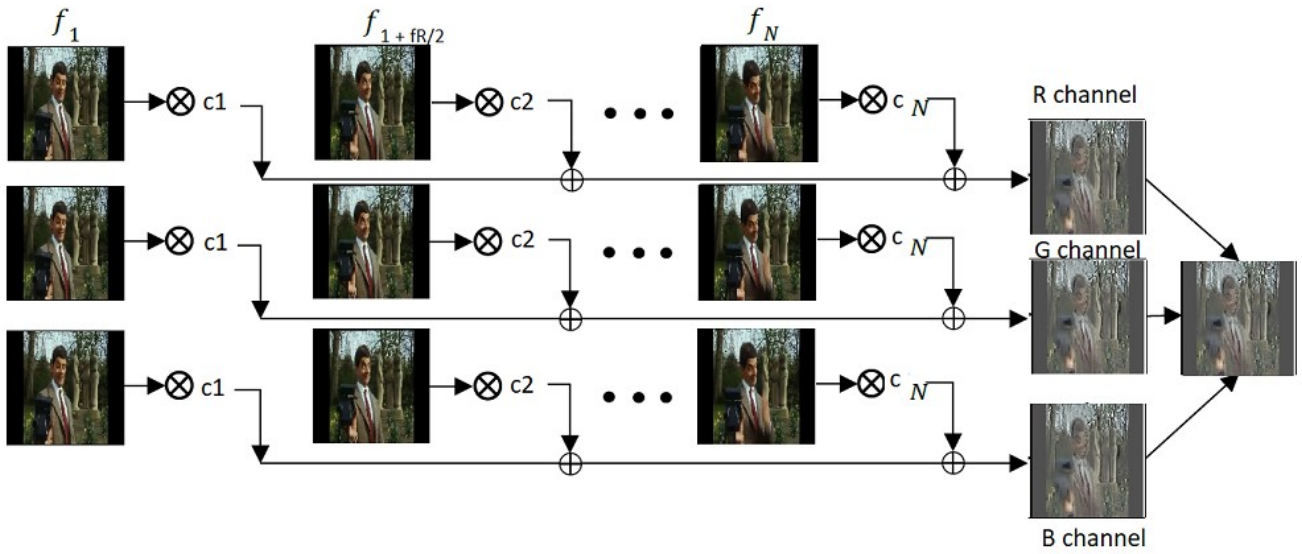

Fig. 3: Dynamic Image Formation

The size of the generated DI was the same as the original frame. Once DI were obtained for the training sample, motion features were extracted from DI using pre-trained Inception-Resnet-V2 on ImageNet dataset followed by fine-tuning layers using Adam optimizer with epoch = 50, and batch size = 64. Callbacks were used to save the best model weights based on minimum validation loss and early stopping with patience = 20 was used to stop the model in case model weights were not improving, hence accelerating the training process. Training input was further divided into actual training set and validation set using 80-20 splitting strategy. In Testing, the model checkpoint saved during training is loaded and DI is computed for each testing video simultaneously. The performance of our proposed architecture on all 3 benchmarks is compiled in Table I highlighting the highest training and validation accuracy achieved and the loss incurred during the training phase as well as the highest testing accuracy achieved during the testing phase.

Table I. Training, Validation, and Testing Accuracy along-with Training, Validation Loss achieved on the 3 publically available benchmarks for our proposed model

Dataset Name	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	Testing Accuracy
Hockey Fight Dataset	100 %	0.0212 %	94.99 %	17.18 %	93.33 %
Real-Life Violence Dataset	100 %	0.0252 %	92.8571 %	21.58 %	86.7892 %
Movie Dataset	100 %	7.30e-03 %	100 %	5.75e-05 %	100 %

Table II. Violence Detection Accuracy (%) Comparison on Hockey Fight Dataset

Method	Accuracy
ViF [22]	81.6%
OViF [24]	84.2%
ViF + OViF [24]	86.03%
HOMO	89.3%
Our Proposed Approach	93.33%

Table II also shows the comparison between the accuracy achieved by our proposed framework and similar works based on motion feature extraction on the Hockey Fight dataset and it is evident that our proposed model has outperformed the rest.

Table III. Violence Detection Accuracy (%) Comparison on Movie Dataset

Method/Classifier		Feature Extracted	Accuracy
[32]		Spatiotemporal	100%
[33]		Spatiotemporal	99.9%
ViF [38]	SVM	Motion	96.7%
	Adaboost		92.8%
	Random Forest		88.9%

Our Proposed Approach	Motion	100%

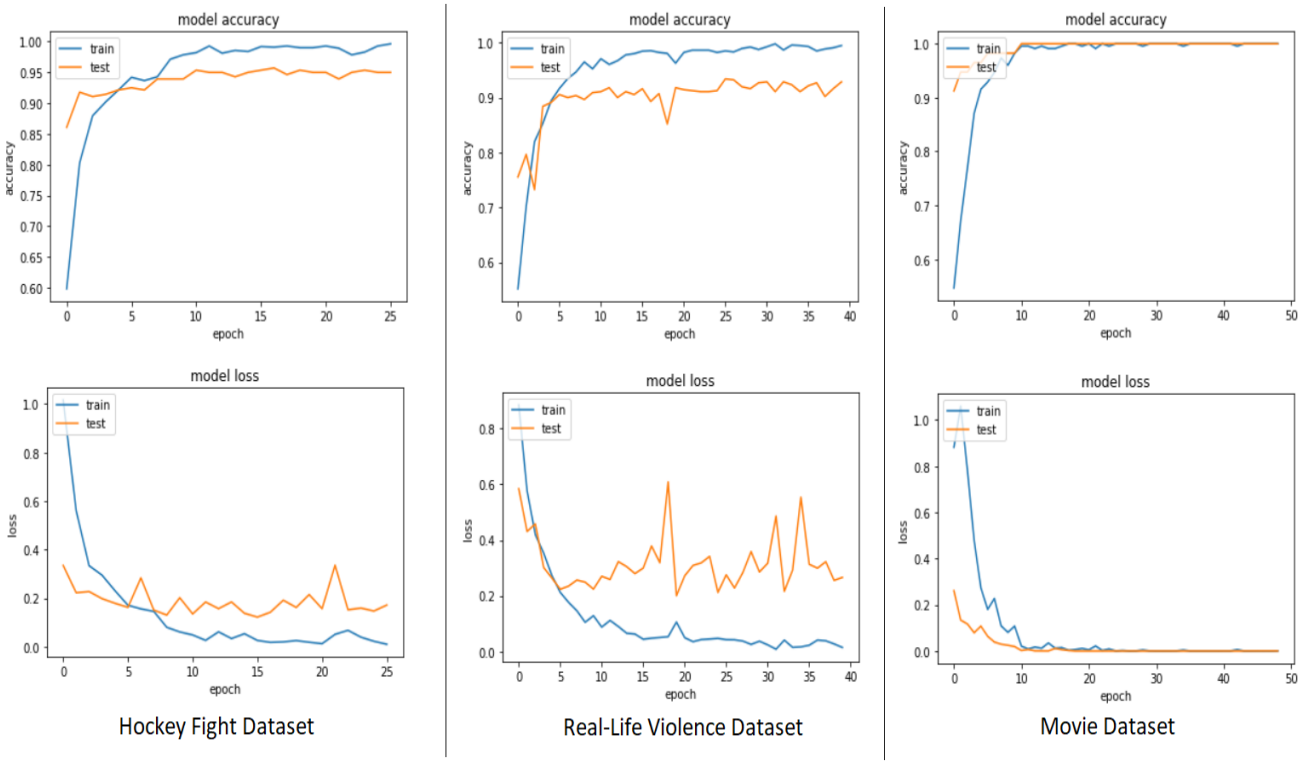

Fig. 4: Plot of Training-Validation Accuracy and Training-Validation Loss during each iteration in the training phase for all 3 dataset

It is also observed that our proposed framework is at par in terms of accuracy achieved with methods based on spatiotemporal feature extraction for detecting violence. Table III shows the accuracy comparison between our framework based on the extraction of motion features and other methodologies based on either extraction of motion or spatiotemporal feature on the movie dataset. The plot of training and validation accuracy, as well as loss incurred on training and validation sample after each iteration in the training phase for all 3 datasets, are shown in Fig 4. On average, it took nearly 1 second or less to convert single RGB video into RGB DI and predict the label as violent or non-violent.

It is also worth noting that even though training accuracy reaches 100% mark, training loss is yet a non-zero value. The reason behind this behavior is that accuracy and loss often appear to be inversely proportional but there is no concrete mathematical relationship between these two metrics [39]. Accuracy can be seen as a count of correct predictions whereas loss can be seen as a distance between predicted probability and true value. The Real-Life Violence Dataset being quite new to the domain isn't used extensively yet, the best testing accuracy achieved by [37] was 88.2% which is higher than achieved by our proposed model. Inception-Resnet-V2 was chosen over other common pre-trained models due to its better top-1 and top-5 accuracy [40].

V. CONCLUSION

With the rising population and increasing need for surveillance has posed growing demands of systems that are capable to detect violent acts automatically. Violence detection is already an interesting research field and CNN's have made an exceptional breakthrough in the field of detecting violence. Although, a great deal of success has been achieved by CNN's yet there a lot of issues that require further investigation. In this work, a novel deep architecture focused on single-stream RGB-DI is introduced that capitalizes on motion features obtained from violent or non-violent videos. DIs are simple yet extremely powerful representations of videos summarizing videos into a single image. DIs can encrypt the summary of the video in a very compact environment allowing for excellent performance by using motion features. Any existing or future CNN architectures can also utilize the DI as input to achieve better results. Furthermore, incorporating DIs in motion stream used in multi-stream state-of-the-arts can enable in learning meaningful results as well. Applying DIs with very recent very deep CNN – pre-trained InceptionResnetV2 with fine-tuning enabled in achieving more speed and better accuracy as compared to existing violence detection methods based on motion feature extraction. Experiments on publically available violence detection dataset benchmarks are conducted to validate the performance and experiments demonstrate the effectiveness of DIs in achieving impressive performance despite their simplicity. In the future, spatiotemporal features of video is aimed to extract along with motion features in a multi-stream architecture to make violence detection more

robust enhancing the performance of the system.

REFERENCES

[1] "New Delhi Streets Turn Into Battleground, Hindus vs. Muslims - The New York Times." [Online]. Available: https://www.nytimes.com/2020/02/25/world/asia/new-delhi-hindu-muslim-violence.html. [Accessed: 21-Jun-2020].

[2] "Fiery Clashes Erupt Between Police and Protesters Over George Floyd Death - The New York Times." [Online]. Available: https://www.nytimes.com/2020/05/30/us/minneapolis-floyd-protests.html. [Accessed: 21-Jun-2020].

[3] B. Jiang, F. Xu, W. Tu, and C. Yang, "Channel-wise Attention in 3D Convolutional Networks for Violence Detection," *Proc. - 2019 Int. Conf. Intell. Comput. Its Emerg. Appl. ICEA 2019*, pp. 59–64, 2019.

[4] E. Ditsanthia, L. Pipanmaekaporn, and S. Kamonsantiroj, "Video Representation Learning for CCTV-Based Violence Detection," *TIMES-iCON 2018 - 3rd Technol. Innov. Manag. Eng. Sci. Int. Conf.*, pp. 1–5, 2019.

[5] M. Baba, V. Gui, C. Cernazanu, and D. Pescaru, "A sensor network approach for violence detection in smart cities using deep learning," *Sensors (Switzerland)*, vol. 19, no. 7, pp. 1–17, 2019.

[6] A. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis, "Bidirectional convolutional LSTM for the detection of violence in videos," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11130 LNCS, pp. 280–295, 2019.

[7] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[9] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018.

[10] J. Mahmoodi and A. Salajeghe, "A classification method based on optical flow for violence detection," *Expert Syst. Appl.*, vol. 127, pp. 121–127, 2019.

[11] K. Lloyd, P. L. Rosin, D. Marshall, and S. C. Moore, "Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures," *Mach. Vis. Appl.*, vol. 28, no. 3–4, pp. 361–371, 2017.

[12] A. A. Einstein, "DETECTION OF REAL-WORLD FIGHTS IN SURVEILLANCE VIDEOS Mauricio Perez , Alex C . Kot School of Electrical and Electronic Engineering University of Campinas Institute of Computing," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 2662–2666, 2019.

[13] A. Singh, D. Patil, and S. N. Omkar, "Eye in the sky: Real-time drone surveillance system (DSS) for violent individuals identification using scatternet hybrid deep learning network," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2018-June, pp. 1710–1718, 2018.

[14] Z. Guo, F. Wu, H. Chen, J. Yuan, and C. Cai, "Pedestrian violence detection based on optical flow energy characteristics," *2017 4th Int. Conf. Syst. Informatics, ICSAI 2017*, vol. 2018-Janua, no. Icsai, pp. 1261–1265, 2017.

[15] A. M. R. Abdali and R. F. Al-Tuma, "Robust Real-Time Violence Detection in Video Using CNN and LSTM," *SCCS 2019 - 2019 2nd Sci. Conf. Comput. Sci.*, pp. 104–108, 2019.

[16] X. Yang, X. Yang, M. Y. Liu, F. Xiao, L. S. Davis, and J. Kautz, "Step: Spatio-temporal progressive learning for video action detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 264–272, 2019.

[17] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, no. i, pp. 1933–1941, Apr. 2016.

[18] T. Z. Ehsan and M. Nahvi, "Violence detection in indoor surveillance cameras using motion trajectory and differential histogram of optical flow," *2018 8th Int. Conf. Comput. Knowl. Eng. ICCKE 2018*, no. Iccke, pp. 153–158, 2018.

[19] L. Lazaridis, A. Dimou, and P. Daras, "Abnormal behavior detection in crowded scenes using density heatmaps and optical flow," *Eur. Signal Process. Conf.*, vol. 2018-Septe, pp. 2060–2064, 2018.

[20] P. D. Garje, M. S. Nagmode, and K. C. Davakhar, "Optical Flow Based Violence Detection in Video Surveillance," *2018 Int. Conf. Adv. Commun. Comput. Technol. ICACCT 2018*, pp. 208–212, 2018.

[21] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3169–3176, 2011.

[22] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 1–6, 2012.

[23] P. Vashistha, C. Bhatnagar, and M. A. Khan, "An architecture to identify violence in video surveillance system using ViF and LBP," *Proc. 4th IEEE Int. Conf. Recent Adv. Inf. Technol. RAIT 2018*, pp. 1–6, 2018.

[24] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using Oriented VIolent Flows," *Image Vis. Comput.*, vol. 48–49, pp. 37–41, 2016.

[25] A. A. Mishra and G. Srinivasa, "Automated detection of fighting styles using localized action features," *Proc. 2nd Int. Conf. Inven. Syst. Control. ICISC 2018*, no. Icisc, pp. 1385–1389, 2018.

[26] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action Recognition with Dynamic Image Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2799–2813, 2018.

[27] C. Dhiman, M. Ieee, D. K. Vishwakarma, and S. Member, "View-invariant Deep Architecture for Human Action Recognition using late fusion."

[28] C. Dhiman and D. K. Vishwakarma, "View-Invariant Deep Architecture for Human Action Recognition Using Two-Stream Motion and Shape Temporal Dynamics," *IEEE Trans. Image Process.*, vol. 29, no. DI, pp. 3835–3844, 2020.

[29] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 6479–6488, 2018.

[30] S. R. Dinesh Jackson *et al.*, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," *Comput. Networks*, vol. 151, pp. 191–200, Mar. 2019.

[31] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," *2017 14th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2017*, 2017.

[32] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," *2017 14th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2017*, 2017.

[33] 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 (Switzerland)*, vol. 19, no. 11, pp. 1–15, 2019.

[34] B. Peixoto, B. Lavi, J. P. Pereira Martin, S. Avila, Z. Dias, and A. Rocha, "Toward Subjective Violence Detection in Videos," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 8276–8280, 2019.

[35] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic Image Networks for Action Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 3034–3042, 2016.

[36] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6855 LNCS, no. PART 2, pp. 332–339, 2011.

[37] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence Recognition from Videos using Deep Learning Techniques," *Proc. - 2019 IEEE 9th Int. Conf. Intell. Comput. Inf. Syst. ICICIS 2019*, pp. 80–85, 2019.

[38] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T. K. Kim, "Fast fight detection," *PLoS One*, vol. 10, no. 4, pp. 1–19, 2015.

[39] "Accuracy and Loss - AI Wiki." [Online]. Available: https://docs.paperspace.com/machine-learning/wiki/accuracy-and-loss. [Accessed: 03-Jul-2020].

[40] Keras Documentation, "Keras Applications," 2017. [Online]. Available: https://keras.io/api/applications/. [Accessed: 03-Jul-2020].