Optical Flow-Based Feature Selection with Mosaicking and FrIFrO Inception V3 Algorithm for Video Violence Detection

Elakiya Vijayakumar
Department of Computer Science and Engineering, Annamalai University, India
elakiyaloganathan@gmail.com (corresponding author)

Aruna Puviarasan
Department of Computer Science and Engineering, Annamalai University, India
arunapuvi95@gmail.com

Puviarasan Natarajan
Department of Computer and Information Science, Annamalai University, India
npuvi2410@yahoo.in

Suresh Kumar Ramu Ganesan
Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, India
aargeek@gmail.com

Received: 16 March 2024 | Revised: 9 April 2024 | Accepted: 14 April 2024

 DOI: https://doi.org/10.48084/etasr.7270

ABSTRACT

Violence in recent years poses the biggest threat to society, which needs to be addressed by all means. Video-based Violence detection is very tough to discern when the person or things that are recipients of a violent act are in motion. Detection of violence in video content is a critical task with applications spanning security surveillance, content moderation, and public safety. Leveraging the power of deep learning, the Violence Guard Freeze-In Freeze-Out Inception V3(VGFrIFrOI3) deep learning model in conjunction with optical flow-based characteristics proposes an effective solution for automated violence detection in videos. This architecture is known for its efficiency and accuracy in image classification tasks and in extracting meaningful features from video frames. By fine-tuning Inception V3 on video datasets annotated for violent and non-violent actions, the network can be permitted to learn discriminative features that simplify the detection of any violent behavior. Furthermore, the aforementioned model incorporates temporal information by processing video frames sequentially and aggregating features across multiple frames using techniques, such as temporal convolutional networks or recurrent neural networks. To assess the performance of this approach, a performance comparison of the proposed model against already existing methods was conducted, demonstrating the model's superior accuracy and robustness in detecting violent actions. The recommended approach not only offers a highly accurate solution for violence detection in video content but also provides insights into the potential of deep learning architectures like Inception V3 in addressing real-world challenges in video analysis and surveillance. The Mosaicking processing, additionally carried out in the pre-processing step, improves the algorithm performance by deploying space search minimization and optical flow-based feature extraction, aiming to extemporize accuracy.

Keywords-deep learning; violence detection; optical flow; convolutional neural networks; InceptionV3; mosaicking

I. INTRODUCTION

The widespread availability of video content has completely changed how information is shared and consumed, with millions of hours of videos being produced and watched every day across a variety of platforms. However, the exponential rise in video content has also brought up new difficulties, most notably the spread of offensive and violent

content. Using deep learning and optical flow-based methods, this study creates a customized CNN model for effective violence detection, which is improved by adding mosaicking. With applications ranging from surveillance to content moderation and video evidence analysis, this system intends to provide real-time, high-precision violence detection by collecting complex motion dynamics and utilizing mosaicking. The results of the investigation could significantly increase detection efficiency. This paper focuses on developing a violence detection method with enhanced feature extraction and classification algorithms so that the latter may be more thoroughly examined. In order to raise the correctness of improvised violence detection, the current study focuses on feature extraction processes in conjunction with classification techniques. Establishing improvised violence detection is not without difficulties. The implementation of conventional violence detection systems is regarded as experimental due to the large amount of data, standards, and noisy images. Computational limitations and visual clarity constitute supplementary problems in violence detection. Applications for video inspection and compression frequently employ the video mosaicking block. These mosaic images are also included with the recovered photos, providing a more complete view of the scene that will help with feature extraction and improve the quality of video categorization. The video mosaicking block is then converted into optical flow frames and given as an input to the training and test phase with the proposed Inception V3 deployed for the detection procedure. Next, a suitable deep learning architecture is chosen for violence detection. Convolutional Neural Networks (CNNs) or pre-trained models like Inception V3 can be effective choices. The dataset is split into training and testing datasets. The selected model is trained using the extracted optical flow frames and corresponding labels (violent or non-violent). The model's hyperparameters, such as learning rate, batch size, and number of epochs, are fine-tuned to optimize performance. The model's performance is analyzed in terms of precision, recall, accuracy, and F1-score once it has been trained on the set used for validation. Once the model performs satisfactorily on the testing set, its performance is assessed by utilizing an untested test set. It is required to continuously monitor the system's performances and gather user feedback to identify the areas of improvement. The model is periodically fine-tuned to adapt to new violence patterns and improve detection accuracy.

The proposed model has various advantages over the existing methods, which rely entirely on hand-crafted features, which have a restricted rate of accuracy and generalization because they are unable to represent the whole complexity and variance in violent behavior. These techniques encounter several difficulties when handling many types of surroundings, namely light flashes, video perspective, and complicated interactions, which may lead to increased rates of false positives and negatives. When it comes to demonstrating notable advances in machine vision tasks, the proposed method employing DL models, like CNN and RNN, has demonstrated notable improvements. Moreover, the long-term dependencies were not accurately detected or captured by the known methodologies. The latter fail to capture time-based correlations that are important for identifying violence. Finally, these cutting-edge methods employ basic classifiers like SVM and RTs. These classifiers do not perform so well because they are less adept at capturing the intricate correlations between data. The proposed approach uses hybrid optical flow-based feature extraction with Inception V3, which combines the freeze in and freeze out concepts, and is inspired by the shortcomings of the current methods.

II. LITERATURE REVIEW

Inception V1 and Inception V3 can be effectively utilized to authenticate people inside an organization deploying handwritten signatures. It appears from research findings on low-resolution photos in the GPDS Synthetic Signature Dataset that Inception V3 is capable of performing better than Inception V1 on high-resolution 3D images like those from ImageNet. The higher-resolution receptive fields used in Inception V3 models are said to produce noticeably better recognition results. In [1], Inception V1, a 22-layer deep network, outperformed the 42-layer Inception V3 network for image input with low resolution. Authors in [2] classified skin cancer implementing SVM along with k-NN classifiers and obtained 61% correct diagnosis [2].

Xception [3] is a unique architecture that shares many of the Inception V3's parameters. On the ImageNet data set, Xception performs slightly better in classification than Inception V3, whereas on the JFT dataset, the former performs much better. Since convolutions that are depth-wise separable have features similar to those of Inception modules and are just as simple to use as ordinary convolution layers, they are expected to become a mainstay of CNN architecture design in the years to come. In [4], the updated Inception V3 with a total accuracy that ranged from 84.8% to 97.8% outperformed classifiers such as kNN, ANN, SVM, linear discrimination, and MLP, which were limited to a total accuracy spanning from 36.4% to 64.6%, dependent on the total number of the visuals utilized for training the classifiers[4]. Transfer learning is a deep learning method, which attempts to solve the image categorization problem. The key benefits of transfer learning approaches are it does not require a large training dataset or a lot of processing capacity [5].

For a small sample, training a full DCNN using a random start is extremely computationally demanding and might not be realistically possible. The pretrained ConvNet can be deployed as an initialization or static extractor of features thanks to transfer learning. Typically, models are trained engaging ImageNet, which comprises over 1.2 million photos classified into 1000 distinct categories [6]. The unprocessed frames technique can discern explosion visual signals and their notable differences in expansion and impact more clearly than their optical path flow or acceleration. Concerning fights, optical movement remained the more accurate descriptor. Given that optical acceleration denotes abrupt pixel changes among frames, the optical flow itself may be a better representation of a fight in a comparatively slower tempo. However, in [7], the optical stimulation provided a more accurate classification of the whole concept of violence. Violence was classified by combining the best outcomes for every notion using fusion networks. Features were perceived at the final fully-connected

layer of the network architecture (Inception), which produced the greatest results, and were fed into the fusion network [7].

The usefulness of utilizing Kinetics initial training for additional video tasks like semantic video categorization, video detection of objects, or optical flow computations is still up for debate [8]. When violent outbursts occur in crowds, prompt discovery could be the difference between life and death. This duty is important, but in the past, it did not receive much attention. In [9], the proposed ViF surpasses the previous methods despite only using the magnitudes of the optical-flow fields [9]. According to [10], bi-LSTMs perform better than ordinary LSTMs in action recognition. Additionally, the attention layer enhances the sequential learning performance. The experimental findings also demonstrate that categorizing fight situations becomes more difficult as the variety a dataset contains increases.

In [11], the motion segmentation as well as the recognition of action experiments showcase the reliability of the predicted optical flow employing FlowNet 2.0 over a wide range of settings and applications. Although the Middlebury findings reveal subpixel motion performance problems, the FlowNet 2.0 results manifest robustness against compression artifacts, retrieving of tiny structures, and extremely precise motion boundaries. In [12], the suggested 3D CNN with an enhanced internal architectural model can efficiently learn the spatiotemporal aspects of aggressive actions with comparatively less parameters. Authors in [13] treated the video sequence as a space-time dimension and utilizing gradients, intensity levels, flow, or other regional properties, the proposed method analyzed activities. Better resistance to deformation, lighting, occlusion, and posture were reported.

In [14], two marine biologists deployed polygons to label the supplied imagery (huge mosaics and several fixed size frames) in an internet-based collaborative annotation process, making sure to achieve pixel-level precision and avoid overlap. Collusion is actually a risky and critical attempt; to eliminate the mark with no lowering the sequence's performance, the succeeding photos are averaged [15].

SURF (Speeded-Up Robust Features) [16], is a popular computer vision technique that is adopted for object detection, picture registration, and classification. It locates and characterizes an image plane's features. It is a more sophisticated and quicker variant of the SIFT technique. Moreover, each segmented area in the image is subjected to a SURF key point-centered approach. This technique is well known for its ability to withstand translation and subsequent processing attacks. In [17], ACO algorithm was employed on the fused features to lower their total dimensional complexity. By keeping the most important characteristics and removing the less significant ones, ACO reduces the complexity of the fused feature vector. To reduce model computation and dimensions, Mobile Net substitutes traditional convolutions by depth-wise convolutions [18].

Deep learning's achievement has ushered in a new phase of artificial intelligence and brought with it fresh approaches to smart video analysis technologies. Large-scale deep learning models and numerous parameters create computation challenges which high-performance GPU chips can address. They also guarantee prompt event detection and analysis. Simultaneously, the substantial quantity of video data satisfies the need for an extensive amount of data instances for deep learning training. The system's data utilization can be effectively enhanced by this type of training [19]. To depict the spatiotemporal video-based actions, in [20], two types of low-level features were proposed to be extracted in the movement regions, Local Histogram of Oriented Gradient (LHOG) and Local Histogram of Optical Flow (LHOF). While LHOF gathers the objects' dynamic information, LHOG may be able to record the appearance of data.

III. THE PROPOSED METHOD

A. Data Collection

In this step, a diverse dataset of video clips containing instances of both violent and non-violent actions is gathered. The acquired hockey dataset encompasses various scenarios, lighting conditions, and camera angles to ensure model robustness. The dataset is an archive of fights between American hockey players, who play in the National Hockey League. It is made up of one thousand video clips, each of which has fifty 360 by 288 frame videos. The video snippets range in duration from 1.6 to 1.96 seconds, with 500 violent and 500 non-violent clips. The information about the data used in the suggested system is displayed in Table I.

TABLE I. DATASET DESCRIPTION

Dataset details	**Hockey fight detection**		
Hockey fight video	20 videos	Hockey no fight videos	20 videos
Hockey fight frame images	20557	Hockey no fight frame images	20500
Training image frames	32846	Testing image frames	8211

B. Optical Flow Frame Extraction and Pre- Processing with Mosaicking

In the preprocessing stage, the optical flow frames should be reseized to a consistent resolution and pixel values should be normalized to be prepared for training. Image mosaicking is a procedure to gather several images from a video to form a single mosaick image, which gives a large view of the scene**.** The pre-processing step of the mosaicking procedure divides the frames into separated subframes having a 35:25 enlargement ratio. The split picture frames are combined during the mosaicking procedure by loading the fight and non-fight datasets. This allows for both feature extraction and categorization to occur. By reducing the amount of search space, the mosaicking method makes use of this segmentation to create image mosaics that improve extraction and classification phase performance and help produce results consistently over time. The image mosaics, or splits of image frames, are progressively matched with both fight and non-fight image frames. After reconstructing the image frame once more, the feature extraction procedure is carried out. Optical flow techniques are adopted to calculate the motion vectors among successive frames in each video. These motion vectors are converted into optical flow frames that capture the dynamic

information of the motions. Optical flow frames provide valuable temporal information that is not present in individual static image frames. In traditional image frames, a snapshot of the scene is captured at a specific moment, and the spatial information alone may not be sufficient enough to infer motion or dynamic changes. However, in the case of videos, the motion of objects over time is crucial for understanding the scene's dynamics, especially in violence detection, where actions and movements are essential indicators.

Optical flow is a computer vision technique that estimates the motion of pixels among successive video frames. It analyzes how the brightness patterns in an image change over time, providing insights into the direction and speed of objects movements. By tracking these motion vectors, optical flow frames are constructed such that the visual flow and dynamics of the scene are captured. The concept behind the optical flow is based on the idea that pixel intensities in a video sequence remain consistent as an object moves. In violence detection, optical flow frames play a vital role in capturing the spatiotemporal features associated with aggressive actions. Actions like punching, kicking, or any violent gestures exhibit specific motion patterns that are recognizable in the optical flow frames. By analyzing the flow of pixels between frames, the violence detection system can learn to identify these patterns, becoming more effective in recognizing violent actions. The use of optical flow frames also allows the model to differentiate between sudden movements characteristic of violence and other regular activities that may involve significant motion, such as running or playing sports. This ability to capture the lively nature of violent actions enhances the model's accuracy and robustness in detecting violence in diverse scenarios. Furthermore, by leveraging optical flow frames, the proposed system can process videos in a more computationally efficient manner compared to analyzing each individual frame independently. Optical flow frames summarize the motion information in the video sequence, reducing the data required for analysis and speeding up the processing time, making it suitable for real-time applications like surveillance.

C. Deep Learning Model Implementation

Inception V3 is a CNN architecture that is widely used for image recognition and classification tasks. In this paper, the Inception V3 model was utilized as the base architecture for training the fight detection model. By utilizing the Inception V3 architecture and fine-tuning it on the fight/non-fight hockey dataset, the trained model learned to abstract more relevant features from the input images and make correct predictions. The utilization of Inception V3 provided a strong foundation for the fight detection model, allowing it to capture complex patterns and spatial dependencies in the image data. Figure 1 depicts the entire structure of the anticipated model.

IV. WORKFLOW OF THE VIOLENCE DETECTION SYSTEM

Inception V3 is known for its efficiency and accuracy in image classification tasks, particularly in large-scale visual recognition challenges like the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Inception V3 is designed to efficiently process and classify images while minimizing computational resources. Tasks involving picture recognition and categorization make extensive use of it. The violence detection technique used in the present work was trained using the Inception V3 model (Figure 2).

The "get_model" function creates a fight detection model by extending the Inception V3 architecture with additional layers. The base model extracts features from the input images and the added layers perform classification based on those features. This model can be trained on fight/non-fight data to learn to classify input images accurately. After defining the "get_generators" function, the "get_model" function is called to obtain the model. Then, the "get_generators" function is called to get the data generators for training and validation. These generators can be employed during the model training process.

The purpose of "freeze_all_but_top" and "freeze_all_but_mid_and_top" functions, is to selectively freeze and unfreeze layers in the model during the fine-tuning or transfer learning. The "freeze_all_but_top" function is utilized to train only the top layers of the model, while keeping the rest of the layers frozen. It takes the "model" as an input. It loops through all the layers in the model except the last two layers (the top layers) and sets their "trainable" attribute to False, effectively freezing those layers. After freezing the desired layers, the model is compiled with the optimizer set to rmsprop, loss function set to "categorical_crossentropy" and accuracy as the evaluation metric. The function returns the updated model. The "freeze_all_but_mid_and_top" function is used to fine-tune the model by selectively freezing and unfreezing specific sets of layers. It takes the "model" as input. It freezes the first 172 layers of the model, including the top layers, by setting their `trainable` attribute to "False". It unfreezes the remaining layers (from the 173rd layer onwards) by setting their "trainable" attribute to True. After modifying the trainable status of the layers, the model is recompiled with the optimizer set to SGD (Stochastic Gradient Descent) with a low learning rate of 0.0001, loss function set to "categorical_crossentropy" and additional metrics for evaluation. The function returns the updated model. These functions are useful when it is needed to fine-tune a pre-trained model by training only specific layers or groups of layers. By freezing some layers, their pre-trained weights are intact and the training process focuses on the remaining layers. This can be helpful when there are limited training data or a need to avoid overfitting. The choice of which layers to freeze or unfreeze depends on the specific task and the desired level of customization. The change from resizing frames to (224, 224) instead of (299, 299) and normalizing them has likely improved the performance and clarity of the model predictions. An explanation of why this change might have made a positive impact is that by normalizing the frames, it is ensured that the pixel values are consistent across different frames and avoid introducing unnecessary variations due to differences in pixel intensity. Normalization can improve the convergence speed and stability of the training process. Finally, the trained model is tested with the test dataset and the proposed VGFrIFrO I3 (Inception V3 Classifier) is applied to predict the fight and no-fight video sequences. Figure 3 presents the entire output screenshots of the entire model.

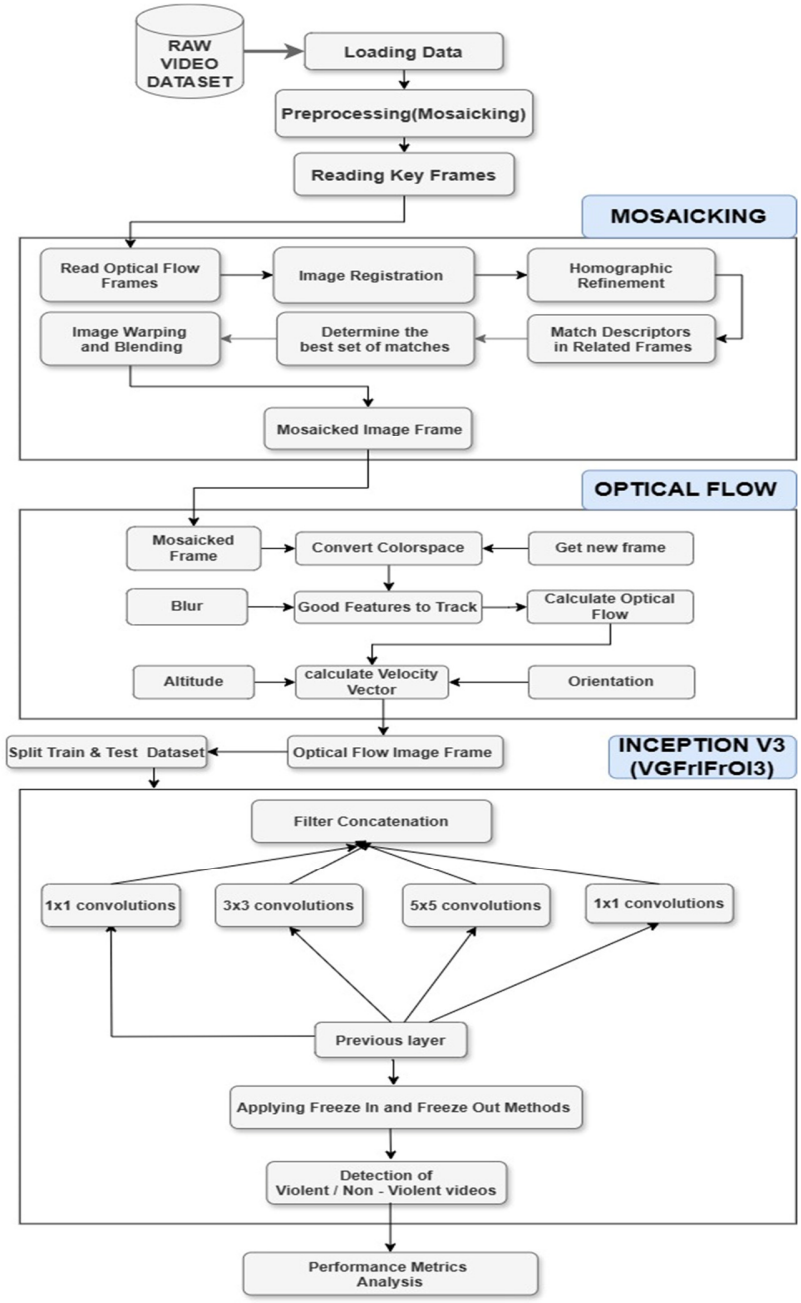

Fig. 1. The proposed VGFrIFrOI3 architecture.

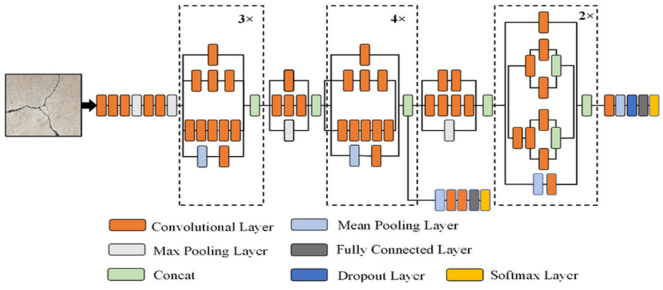

Fig. 2. Inception V3 architecture.

V. RESULTS

A. Performance Assessment Metrics

In order to improve the method's ability to distinguish between fight and non-fight frames, the procedure for training entails identifying frames that are frozen and unfrozen. Iterative steps are taken to train the framework until it reaches the point of accurate detection. Using algorithms to find potential matches to map the best inputs to the output is the process of training the model. The model is trained by an iterative approach that updates the model's freeze in and freeze out layers in each iteration.

The efficiency analysis of the proposed violence detection model employing VGFrIFrOI3 is demonstrated implementing various metrics. In the following TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative.

1) Accuracy

One of the easiest classification measures to use is accuracy, which can be calculated as the ratio of accurate predictions compared with all predictions.

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP}$$

- Precision

The percentage of positive predictions that were truly accurate is determined by the precision metric.

$$\text{Precision} = \frac{TP}{FP+TP}$$

- Recall

The purpose of this step is to figure out which percentages of real positives were misidentified. Either accurately forecasted as positive or erroneously projected as negative, they can be measured as true positive or as forecasts that are true to the entire number of positives.

$$\text{Recall} = \frac{TP}{FN+TP}$$

Fig. 3. Output screenshots.

A test set of 20% and a training set of 80% of the mosaicked image frames in the hockey conflict dataset were considered. Different traditional methods using different datasets determine different values of accuracy. Based on the results of Table I, it was found that the proposed method for classifying fight and no-fight (violent and non-violent) images from hockey collection clips of the video produces scores an accuracy of 98.5%, surpassing the conventional techniques. Tables II and III and Figures 4 and 5 display the accuracy of the proposed model.

TABLE II. MODEL ACCURACY.

Model	Accuracy
ResNet	84.09%
Inception V3	87.42%
KNN	64.6%
MobileNet with Transfer Learning	91%
Xception+LSTM	97.5%
Proposed Method VGFrIFrOI3	98.5%

TABLE III. DATABASE ACCURACY

Model	Accuracy
ResNet	84.29%
Inception V3	89.42%
KNN	80.6%
MobileNet with Transfer Learning	91%
Xception+LSTM	92.5%
Proposed Method VGFrIFrOI3	95.9%

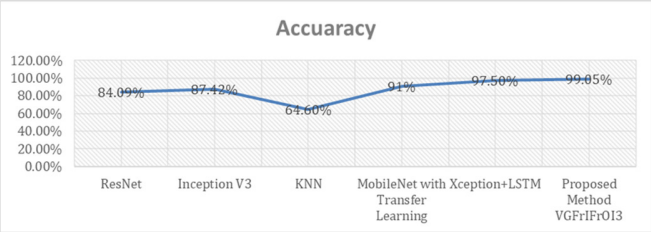

Fig. 4. Accuracy graph.

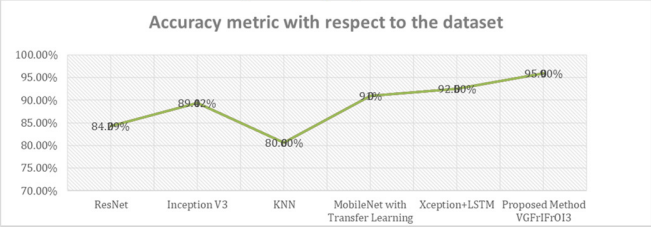

Fig. 5. Dataset accuracy metrics with diverse classifiers.

With regard to the identification of fight sequences, the classification results of the proposed system using the VGFrIFrOI3 achieved model greater accuracy (98.5%), poor precision (0.97%), increased recall (0.97%), and short specificity score (97%) in comparison with those of the conventional approaches.

VI. CONCLUSION

The results section presents the condensed performance analysis and comparison evaluation of the optical flow extraction of the feature model outcomes. The mosaicked frames with the best matches among all frames and at the end the series of mosaicking process gives the best patched frames, which adds to the performance at the preprocessing step. In identifying whether an image frame contains a fight scene or not, the proposed VGFrIFrOI3 violent detection system gives a high accuracy of 98.5%, sensitivity of 96.4%, and specificity of 95.2%, outperforming the other conventional violence detection systems by using diverse feature abstraction and classifiers that are traced from numerous datasets. With a similar vein, the proposed model also demonstrated improved performance results in terms of the score values of accuracy and recall parameters when deployed to identify violent frames from the hockey dataset, explaining the increased effectiveness of the proposed Model.

VII. FUTURE SCOPE

Even though a lot of research has been done on automatic violence-activity recognition in the past ten years, automatically recognizing and comprehending human behavior and activity is still a laborious task, that faces issues including occlusion intersect, lightning and brightness variations, camera movement, multiple camera views, and real-world dynamics of photographic camera resolution. Other problems that prevented systems from identifying the behavior of image frames include the absence of labelled data, fluctuations in similar activity, distinct patterns of violent as well as human behavior and complex noise. Distinct answers should be offered to each of those difficulties separately, so that those issues could be addressed successfully and prominently.

REFERENCES

[1] Jahandad, S. M. Sam, K. Kamardin, N. N. Amir Sjarif, and N. Mohamed, "Offline Signature Verification using Deep Learning Convolutional Neural Network (CNN) Architectures GoogLeNet Inception-v1 and Inception-v3," *Procedia Computer Science*, vol. 161, pp. 475–483, Jan. 2019, https://doi.org/10.1016/j.procs.2019.11.147.

[2] A. Demir, F. Yilmaz, and O. Kose, "Early detection of skin cancer using deep learning architectures: resnet-101 and inception-v3," in *2019 Medical Technologies Congress (TIPTEKNO)*, Izmir, Turkey, Oct. 2019, pp. 1–4, https://doi.org/10.1109/TIPTEKNO47231.2019.8972045.

[3] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 1800–1807, https://doi.org/10.1109/CVPR.2017.195.

[4] A. E. Tio, "Face shape classification using Inception v3." arXiv, Nov. 14, 2019, https://doi.org/10.48550/arXiv.1911.07916.

[5] M. M. Rahman, A. A. Biswas, A. Rajbongshi, and A. Majumder, "Recognition of Local Birds of Bangladesh using MobileNet and Inception-v3," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 8, 31 2020, https://doi.org/10.14569/IJACSA.2020.0110840.

[6] N. Aneja and S. Aneja, "Transfer Learning using CNN for Handwritten Devanagari Character Recognition," in *2019 1st International Conference on Advances in Information Technology (ICAIT)*, Chikmagalur, India, Jul. 2019, pp. 293–296, https://doi.org/10.1109/ICAIT47043.2019.8987286.

[7] B. Peixoto, B. Lavi, P. Bestagini, Z. Dias, and A. Rocha, "Multimodal Violence Detection in Videos," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, Feb. 2020, pp. 2957–2961, https://doi.org/10.1109/ICASSP40776.2020.9054018.

[8] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 4724–4733, https://doi.org/10.1109/CVPR.2017.502.

[9] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, Jun. 2012, pp. 1–6, https://doi.org/10.1109/CVPRW.2012.6239348.

[10] Ş. Aktı, G. A. Tataroğlu, and H. K. Ekenel, "Vision-based Fight Detection from Surveillance Cameras," in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Istanbul, Turkey, Nov. 2019, pp. 1–6, https://doi.org/10.1109/IPTA.2019.8936070.

[11] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1647–1655, https://doi.org/10.1109/CVPR.2017.179.

[12] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient Violence Detection Using 3D Convolutional Neural Networks," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Taipei, Taiwan, Sep. 2019, pp. 1–8, https://doi.org/10.1109/AVSS.2019.8909883.

[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*, 2011, pp. 332–339, https://doi.org/10.1007/978-3-642-23678-5_39.

[14] K. Buskus, E. Vaiciukynas, S. Medelytė, and A. Šiaulys, "Exploring the necessity of mosaicking for underwater imagery semantic segmentation using deep learning," *Journal of WSCG*, vol. 30, no. 1–2, pp. 26–33, Jan. 2022, https://doi.org/10.24132/JWSCG.2022.4.

[15] S. Mansour, S. Ben Jabra, and E. Zagrouba, "A Robust Deep Learning-Based Video Watermarking Using Mosaic Generation:," in *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Lisbon, Portugal, 2023, pp. 668–675, https://doi.org/10.5220/0011691700003417.

[16] U. Diaa, "A Deep Learning Model to Inspect Image Forgery on SURF Keypoints of SLIC Segmented Regions," *Engineering, Technology & Applied Science Research, vol. 14, no. 1, pp. 12549–12555, Feb. 2024, https://doi.org/10.48084/etasr.6622.*

[17] T. Imran, A. S. Alghamdi, and M. S. Alkatheiri, "Enhanced Skin Cancer Classification using Deep Learning and Nature-based Feature Optimization," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12702–12710, Feb. 2024, https://doi.org/10.48084/etasr.6604.

[18] V. A. Rajendran and S. Shanmugam, "Automated Skin Cancer Detection and Classification using Cat Swarm Optimization with a Deep Learning Model," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12734–12739, Feb. 2024, https://doi.org/10.48084/etasr.6681.

[19] P. Wang, P. Wang, and E. Fan, "Violence detection and face recognition based on deep learning," *Pattern Recognition Letters*, vol. 142, pp. 20–24, Feb. 2021, https://doi.org/10.1016/j.patrec.2020.11.018.

[20] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence detection in surveillance video using low-level features," *PLOS ONE*, vol. 13, no. 10, 2018, Art. no. e0203668, https://doi.org/10.1371/journal.pone.0203668.