A deep Spatio-temporal network for vision-based sexual harassment detection

Md Shamimul Islam
Dept. of CSE
Manarat International University
Dhaka, Bangladesh
shamimulislam@manarat.ac.bd

Md Mahedi Hasan
IICT
BUET
Dhaka, Bangladesh
mahedi0803@gmail.com

Sohaib Abdullah
Dept. of CSE
Manarat International University
Dhaka, Bangladesh
sohaib@manarat.ac.bd

Jalal Uddin Md Akbar
Dept. of CSE
International Islamic University Chittagong
Chittagong, Bangladesh
jalaluddinmdakbar00@gmail.com

N H M Arafat
Dept. of Computer Science and Technology
Henan Polytechnic University
454003, Jiaozuo, Henan, P.R. China
arafat.nhm@gmail.com

Saydul Akbar Murad
Faculty of Computing
Universiti Malaysia Pahang
26600, Pahang, Malaysia
saydulakbarmurad@gmail.com

2021 Emerging Technology in Computing, Communication and Electronics (ETCCE) | 978-1-6654-8364-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ETCCE54784.2021.9689891

***Abstract*—Smart surveillance systems can play a significant role in detecting sexual harassment in real-time for law enforcement which can reduce the sexual harassment activities. Real-time detecting of sexual harassment from video is a complex computer vision because of various factors such as clothing or carrying variation, illumination variation, partial occlusion, low resolution, view angle variation etc. Due to the advancement of convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM), human action recognition tasks have achieved great success in recent years. But sexual harassment detection is addressed due to presences of large-scale harassment dataset. In this work, to address this problem, we build a video dataset of sexual harassment, namely Sexual harassment video (SHV) dataset which consists of harassment and non-harassment videos collected from YouTube. Besides, we build a CNN-LSTM network to detect the sexual harassment in which CNN and RNN are employed for extracting spatial features and temporal features, respectively. State-of-the-art pretrained models are also employed as a spatial feature extractor with an LSTM and three dense layer to classify harassment activities. Moreover, to find the robustness of our proposed model, we have conducted several experiments with our proposed method on two other benchmark datasets, such as Hockey Fight dataset and Movie Violence dataset and achieved state-of-the-art accuracy.**

***Index Terms*—sexual harassment, surveillance systems, deep learning**

I. Introduction

Sexual harassment, one of the most unwanted and intolerable criminal activities in the world, which according to the UN, is unsolicited sexual advances, asking for sexual favors, and other physical or verbal activities which are sexual in nature [1]. As it is accomplished against one's will, it results in hurting the victim's dignity. In a variety of circumstances, it may occur in different places such as in workplaces, in the military, in academic institutes, during transportation [2], in public places, etc. The harasser can be a direct supervisor, an indirect supervisor, a coworker, an instructor, a peer, or a colleague and may have any gender and any type of connection with the victim. It can happen in physical or verbal form and both online and offline. In this paper, our focus is on the onsite physical form of harassment as it is one of the most serious offenses and if not protected in due time may lead to greater crimes like rape.

Nearly one in five women will experience sexual harassment in any given year, according to the result of a survey named "National Intimate Partner and Sexual Violence Survey" [3]. Victims of sexual harassment can experience severe psychological consequences, including anxiety, headaches, depression, sleep disorders, losing or gaining weight, nausea, reduced self-esteem, and sexual dysfunction. Using advanced intelligent surveillance-based systems, we can reduce this unwanted event. Although sexual harassment cannot be completely prevented, the sexual harassment detection system can reduce the frequency, if it can accurately recognize a harassment incident and generate an alert.

In this work, we have explored several deep learning algorithms to classify harassment in videos and experimented with those on the Sexual Harassment Video (SHV) dataset which has been collected from YouTube. The dataset contains harassment videos that have been collected from different dramas, harassment training videos, harassment consciousness videos, etc. We have applied CNN-LSTM in our collected dataset where CNN is employed as a spatial feature extractor and LSTM is used as a temporal feature extractor. We have evaluated our proposed model extensively in two of the most complicated and benchmark datasets, the Hockey fight dataset [6], and Movie Violence Dataset [6]. The proposed method achieved state-of-the-art accuracy in these datasets. The following are the paper's major contributions:

- We have developed a novel Sexual Harassment Video (SHV) dataset which contains 300 sexual harassment videos belonging to two classes: harassment and non-harassment situations. We have collected events that occurred in different locations in such a manner that it can represent both in-house and open site events.

978-1-6654-8364-3/21/31.002021*IEEE*

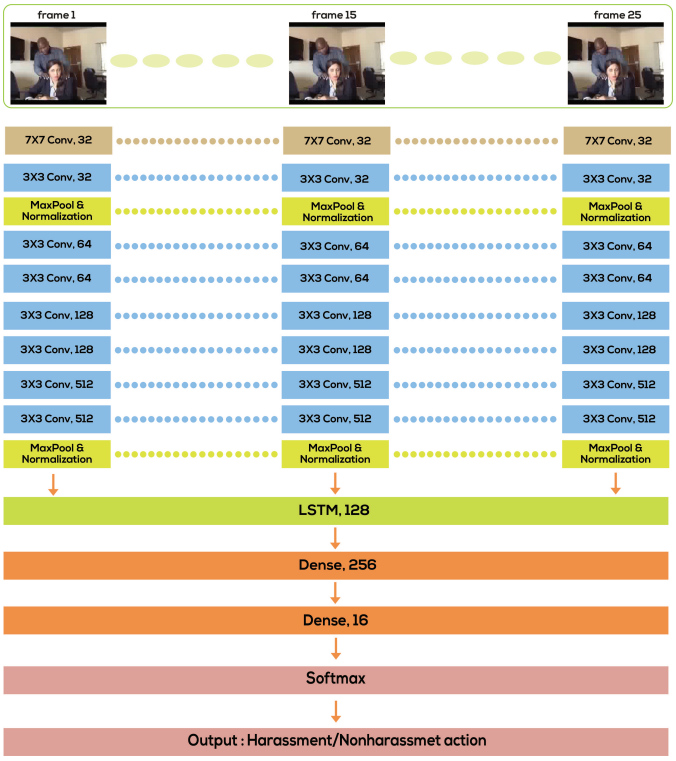

Fig. 1: The network architecture of the proposed model. The model consist of ten CNN layers and one LSTM. CNN is used to extract spatial features and LSTM is employed to extract temporal features. Lastly, a softmax layer is employed to classify the harassment and non-harassment activity.

- We have built a deep learning-based model combining CNN and LSTM to classify the harassment and non-harassment events.
- We used two benchmark datasets, namely Hockey fight dataset [4] and Movie Violence Dataset [4], to validate the proposed model's performance and found it to be state-of-the-art.

II. Related Work

The task of classifying sexual harassment in a given video can be defined as an action recognition problem. Traditionally, in order to recognize action [5], handcrafted feature representation-based approaches were used. Some of such handcrafted representation techniques include spatiotemporal volume-based techniques [6], STIP-based methods [9], motion trajectory-based methods. Deep learning-based methods offer an alternative to feature-based methods nowadays [10]. But traditional methods are still in use basically because of the computational [12] power involved in deep learning approaches. However, action like sexual harassment recognition involves a complex scenario in which hand-crafted feature-based methods are not capable of performing well.

Deep learning-based human activity recognition makes a great stride with the improvement of the GPU and, the compilation of the large-scale training dataset. Two-stream convolutional networks are one of the most popular deep learning approaches for action recognition [13] in which two CNN layers are used and outputs of those layers are mixed at the end. Zhang et al. [15], used the motion vector in the video stream to increase the measurement speed and realize real-time recognition [16] of human activity. A temporal segment network (TSN) has been proposed by Wang et al. [18] to further enhance the performance of the two-stream convolutional network. The recognition accuracy of TSN was also improved by Lan [19] and [20]. Carreira et al. [21] expanded from 2D to 3D the inception-V1 network structure and proposed the two-stream inflated 3-dimensional ConvNet to recognize the action. In [17], [21], authors used another important method combining the LSTM and CNNs for human action recognition. Liu et al. [23] proposed a 3D human spatiotemporal LSTM model for action recognition. For classification, Sudhakaran and Lanz proposed convolutional LSTM in order to better distinguish between Spatio-temporal changes between frames [14].

To the best of our knowledge, there is not that much work available on sexual harassment detection from video. MDT Shahria et al. [7] used a CNN-based model for detecting sexual harassment from video in the workplace. Ng et al. [22] compared the CNN-only model with other combined models. The CNN-based model encodes local temporal features in each individual video frame only (which forms a small part of the action of interest) and makes an average of the predictions. It is incapable of modeling across multiple video frames over a video clip of interest. It bears the risk of missing vital information and creating confusion among classes. So this approach is not that suitable in complex scenarios in which there are fine-grained distinctions. However, if CNN is combined with LSTM where output features extracted from the CNN becomes the input of the LSTM, this combination can become efficacious in the representation of long-term motion and model across sequential video frames representing the motion of interest effectively. In this work, to classify harassment events, LSTM is employed that are connected with the output of a CNN model in order to model global sequence dependencies and motion information more accurately. Morever [7] used a dataset that contains in total of 100 workspace videos whereas this paper works with a dataset that contains three times more data that contains harassment events not only in the workplace but also in different situations such as in transportation, in public places, etc.

III. Methodology

Fig. 1 shows the workflow of the sexual harassment detection model. To classify sexual harassment activity, our proposed model is capable to predict sequences in consecutive frames. In video analysis, the promising result of classification depends on a model that can track the connection between temporal features and spatial features. In this work, at first, we perform several preprocessing steps on selected frames from raw input harassment videos. In order to capture the spatial features from processed frames, we employ a CNN model. After that, the spatial features are extracted from each frame

and a time-step of spatial features is formed to feed into an LSTM-based network to capture the temporal features. Finally, the fully connected layers are used to classify the harassment action.

A. Scratch Model

We propose an advanced deep learning-based Sexual Harassment Detection Network (SHDN) in order to better understand the dataset's insight pattern. SHDN is developed by the combination of CNN and LSTM. At first, the preprocessing actions are performed in frames that are extracted from raw videos. After that, the processed frames of the video dataset are fed into the CNN architecture, which then extracts spatial features from those processed frames. Our proposed CNN architecture of SHDN has ten layers including convolutional layers and pooling layers. The total convolutional layers of our proposed model are eight and the number of filters in those eight convolutional layers are as follows (sequentially from first to last): $32, 32, 64, 64, 128, 128, 512, 512$. Every filter has a size of 3×3 in CNN architecture except the first one which has a size of 7×7. In our model, we use "same-padding". Two max-pooling layers are used in our model. The last layer of our CNN architecture is a flatten layer. A batch normalization [34] is In this study, the dropout rate of 0.2 is employed to the fully connected layer in order to reduce the overfitting problem. Adam [30] optimizer with .00001 learning rate is applied in all models. The epochs number is 50 for each model with batch size of 16.mployed in our network. We use Xavier uniform [36] initializer as kernel initializer. To prevent overfitting problems, we use regularization techniques like kernel regularizer L2. The LeakyReLU [35] is used as an activation function for our model. We use the Time distributed layer of Keras to implement CNN architecture.

After that, we pass the 1D spatial features vector into the LSTM layer of 256 hidden units to extract the temporal features. Then, two dense layers are used in our model which has 512 and 16 neurons, respectively. The dropout rate of the dense layer is 0.2. Finally, we use the output layer of two neurons with the Softmax activation function to classify the sexual harassment action. Adam [34] optimizer is employed for this architecture. Categorical cross-entropy is used as a loss function for our algorithm 1.

$$L\_s = -\sum_{i=1}^{m} log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{1}$$

The total trainable parameters of the proposed model is 4,171,218. The architecture of our proposed model is depicted in Fig. 1.

B. Transfer Learning Approach

A rule of thumb of deep learning is that a lot of data is needed to train a deep learning network from scratch for solving a problem effectively [24]. In reality, such a vast amount of data is not always available. Transfer learning has a better ability to perform greatly with a small dataset [25]. However, transfer learning is employed for extracting spatial features in this work. In recent years, several pre-trained models have achieved quite a popularity. VGG [26], ResNet [27], Xception [28], Inception v3 [31], MobileNet [29] are named of most popular pre-trained model that were trained for the ImageNet "Large Visual Recognition Challenge" [8] images. In this paper, we explored the pre-trained Inception v3 [31], VGG16 [26], VGG19 [26], Xception [28], ResNet50 [27], MobileNet [29] and InceptionResNet [30] to extract spatial features from Sexual harassment video dataset and then pass into LSTM network in order to extract temporal features.

IV. EXPERIMENTAL RESULTS

A. Dataset

In this study, we use three datasets our newly developed Sexual Harassment Video (SHV) dataset, Hockey Fight Dataset [4], Movies Dataset [4], to validate our model's performance.

TABLE I: Sexual Harassment Video dataset details.

Sexual Harassment Video (SHV) Dataset	
Total Classes	2
Number of Videos per class	150
Frames per second	25
Video Length	1-3s
Avg. frames per video	50
Resolution	320×240
No of location	3

1. Sexual Harassment Video (SHV) dataset: SHV dataset is the primary contribution of this paper. SHV dataset is collected from various video hosting sites in the context of physical sexual harassment in indoor and outdoor environments such as streets, markets, workplaces, educational institutes, etc. We have collected videos of different physical sexual harassment such as an unwanted touch on the sensitive organ of the body, molestation, etc., and labeled them as harassment videos. Moreover, we collected videos of peaceful events where normal people are walking on the street, normal transportation views, normal office work, etc., and label them as non-harassment videos. We have collected 150 videos per class. Then, we put 45 videos per class for testing and 105 videos per class for training purposes. The first four images of the first and second row of Fig. 2 depicts some sample frames of the SHV dataset that the first row is harassment and the second row is non-harassment.

2. Hockey Fight Dataset [4]: The Hockey dataset has 1000 video clips with a resolution of 720×576 pixels, categorized as follows, five hundred fights and five hundred non-fights. The frame numbers of each clip are 50 and each video is two seconds long.

3. Movies Dataset [4]: The total number of videos is 200, split into equally two parts. The fight clips part was collected from Hollywood movies and the non-fight part was collected from football and other events. The average frame numbers are

Fig. 2: Sample video frames randomly selected from three Datasets. In first row, the first four images are harassment class from SHV dataset and last four images are violent class from Hockey Fight Dataset [4] and Movies Dataset [4]. In second row, the first four images are non-harassment class from SHV dataset and last four images are non-violent class from Hockey Fight Dataset [4] and Movies Dataset [4].

not equal in the dataset and the average duration is 2second per clip.

TABLE II: Precision, recall & F-1 score values are compared for different architectures. C = CNN layer, R = RNN layer & D = Dense layer. C10R1D2 gives the best performance in our experiments which consist of 10 CNN layers, 1 RNN(LSTM) layer and 3 Dense layers.

Architecture Name	Architecture	Precision	Recall	F-1 Score	AUC
Network-1	C4R1D2	80	74	76.92	80
Network-2	C6R1D2	88	56	68	75
Network-3	C8R1D2	81	67	73	78
Network-4	C10R1D2	91	78	84	86
Network-5	C4R1D3	80	74	76.92	80
Network-6	C6R1D3	88	56	68	75
Network-7	C8R1D3	80	74	76.92	80
Network-8	C10R1D3	87.50	94	90	88

TABLE III: From the table, it has been observed that normalized data gives good performance.

	Testing Accuracy	Total Epochs
Normalized Data	89.06%	50
Unnormalized Data	87.50%	50

TABLE IV: Precision, recall and F-1 score with AUC values are compared for each dataset

Dataset	Values							
	TP	TN	FP	FN	Prec.	Recall	F-1	AUC
SHV Dataset	35	22	2	5	88	94	90	88
Hockey Fight	48	49	2	1	96	98	97	97
Violence in Movies	15	15	0	0	100	100	100	100

B. Implementation details

We use the Keras DL package by TensorFlow to implement the aforementioned CNN-LSTM networks. We conduct our experiment by using the NVIDIA GeForce GTX 1050 GPU. We resize to $128 \times 128 \times 3$ for our network inputs in order to reduce computational complexity. The dataset is divided: 80% for training and 20% for testing. Our proposed CNN-LSTM-based model is trained using Adam [35] optimizing algorithm with 0.0001 learning rate and categorical cross-entropy loss function. The Xavier [36] algorithm is employed as a weight initializer in our model. The batch size for our CNN-LSTM model was 16. L2 regularizer was used as a kernel regulizer to reduce the over-fitting problem. A total of 200 epoch is used during training, each one comprising 325 iterations.

TABLE V: the confusion matrix for three dataset.

SHV Dataset		Hockey Fight [4]		Violence in Movies [4]	
Haras.	Non-haras.	Violent	Non-voilent	Violent	Non-violent
35	2	48	2	15	0
5	22	1	49	0	15

C. Results

In our experiment, as in Table II, we have tried several network architectures which comprise CNN, LSTM, and fully connected layers in order to find the optimal model. Comparison is mentioned in Table II, Network-8 performs better than other networks which consist of ten CNN layers, one LSTM layer, and three dense layers. Network-8 gives 87.50% precision, 94% recall, 90% F1-score and 88% AUC. We try ReLU and LeakyReLU [35] as an activation function for every network in the experiment. LeakyReLU [34] activation function performs better for our dataset. SGD and Adam [34] are used as an optimization algorithm. The best optimization [33] algorithm is Adam [34] for our experiment.

We employ normalization to speed up the model convergence during training. The performance comparison of the optimal model on normalized and unnormalized data is represented in Table III. According to the table III, due to the normalization, better performance has been achieved within fifty epochs compared to the training data that hasn't been normalized.

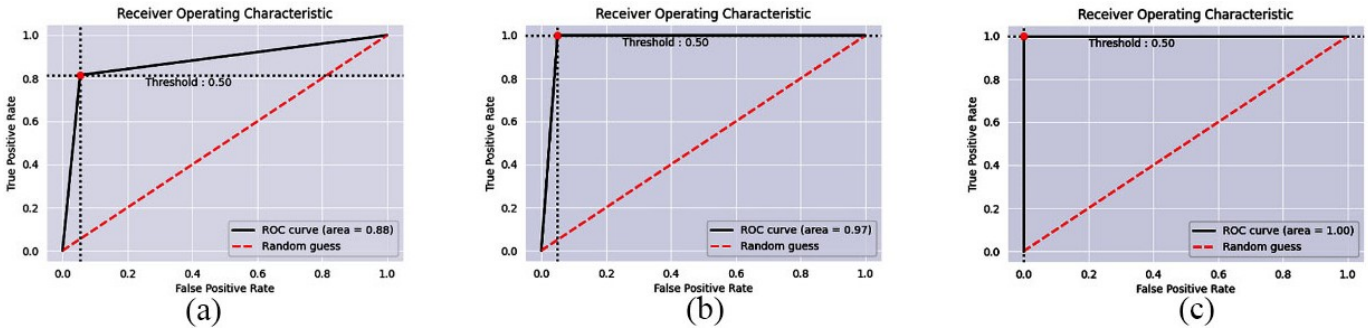

Fig. 3: The ROC curve based on AUC value for (a)Sexual Harassment Video Dataset(SHV);(b)Hockey Fight Dataset [4] and (c)Movies Dataset [4]. Curve closer to the top-left corner gives the best performance. Experiment on Hockey Fight Dataset touches the most top-left corner.

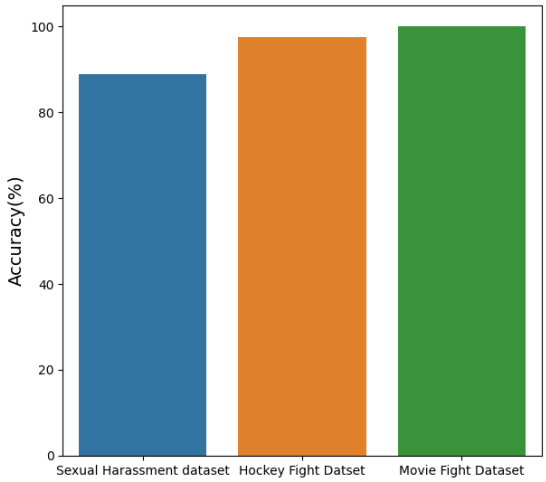

Fig. 4: The comparison of the proposed method on three datasets based on accuracy.

We assess the proposed CNN-LSTM based method's performance by analyzing precision, recall, and the comparability of datasets using the area under the curve values (AUC) in Table IV, which shows the robustness of the proposed method. Additionally, Table V contains the obtained confusion matrix. The precision achieved on SHV dataset, Hockey Fight Dataset [6], Movies Dataset [6] are 96%, 100%, and 98%, respectively, while the recall is 96.67%, 100%, and 98.76%, respectively. We further used the Receiver Operating Characteristic(ROC) curve in order to evaluate the effectiveness of the proposed model in Fig. 3. Here using Area under Curve(AUC) in Fig. 3. AUC value obtained for SHV dataset, hockey fight, violence in movies is 88.00%, 97.00%, 100% respectively. We also compared the accuracies for our proposed dataset SHV and two benchmark datasets in Figure 4 by plotting bar chart, where the highest achieved is 99.9% obtained in the movies dataset, 98% accuracy is obtained in the hockey fight dataset, and 89% accuracy is obtained in our self-developed Sexual harassment dataset.

Additionally, we conduct an experiment using six of the top most popular pre-trained models(Inception v3 [31], VGG16 [26], VGG19 [26], Xception [28], ResNet50 [27], MobileNet [29] and InceptionResNet [30]) in combination with LSTM and dense layer. Table VI depicts the performance comparison of these models on the SHV dataset. In this experiment, transfer learning approaches by using the VGG16 [26] and MobileNet [29] models get better accuracy than the others in classifying SHVD classes.

TABLE VI: Precision, recall, F-1 score and AUC values are compared for SHV dataset using popular pretrained model

Pretrained Models	Precision	Recall	F-1 Score	AUC
VGG16 [26]	90	97	93	91
VGG19 [26]	88	56	68	75
Inception [31]	80	74	77	80
ResNet50 [27]	85	95	90	86
Xception [28]	82	89	85	82
MobileNet [29]	86	97	91	88
InceptionResNet [30]	81	67	73	78

In this study, the dropout rate of 0.2 is employed to the fully connected layer in order to reduce the overfitting problem. Adam [34] optimizer with .00001 learning rate is applied in all models. The epochs number is 50 for each model with batch size of 16.

D. Conclusion

The major objective of this work was to build a model for classifying sexual harassment events in videos while maintaining state-of-the-art accuracy. Our proposed method comprises the CNN and LSTM layers to extract spatial and temporal features for classifying the harassment activities. Moreover, using popular pre-trained models for feature extraction achieved great accuracy in our SHV dataset experiments. Furthermore, our proposed model was also evaluated over two other benchmark datasets: Hockey Fight Dataset, Movies Dataset.

The collection of the Sexual harassment video (SHV) dataset is another contribution to building sexual harassment detection model. In the future, to construct a more reliable network for practical applications such as real-time surveillance-based sexual harassment detectors, we will extend our work

by adding a larger multi-view dataset containing RGB videos of thousands of subjects.

REFERENCES

[1] United Nations, "What is sexual harassment," [Online] Availble: https://www.un.org/womenwatch/osagi/pdf/whatissh.pdf. [Accessed: Nov. 10, 2020]

[2] A. J. Md Muzahid, S. F. Kamarulzaman, and M. A. Rahim, "Learning-based conceptual framework for threat assessment of multiple vehicle collision in autonomous driving," in2020 Emerging Technology in Computing, Communication and Electronics (ETCCE), 2020, pp. 1–6

[3] "Nearly 1 in 5 Women in U.S. Survey Say They Have Been Sexually Assaulted," https://www.nytimes.com/2011/12/15/health/nearly-1-in-5-women-in-us-survey-report-sexual-assault.html, 2011.

[4] E. B. Nievas, O. D. Suarez, G. B. Garcıéa, and R. Suk thankar, "Violence detection in video using computer vision techniques." In International Conference on Computer Analysis of Images and Patterns, Springer, 2011.

[5] Hasan, Md Mahedi, Md Shamimul Islam, and Sohaib Abdullah. "Robust pose-based human fall detection using recurrent neural network." In 2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON), pp. 48-51. IEEE, 2019.

[6] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257–267, March 2001, doi: 10.1109/34.910878.

[7] M. T. Shahria, F. Tasnim Progga, S. Ahmed and A. Arisha, "Application of Neural Networks for Detection of Sexual Harassment in Workspace," 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2021, pp. 1-4, doi: 10.1109/ICAECT49130.2021.9392429.

[8] O. Russakovsky*, J. Deng*, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and Li Fei-Fei,(* = equal contribution) "ImageNet Large Scale Visual Recognition Challenge," IJCV, 2015.

[9] I. Laptev, "On Space-Time Interest Points." Int J Comput Vision 64, pp. 107–123 (2005). https://doi.org/10.1007/s11263-005-1838-7

[10] S. A. Murad, Z. R. M. Azmi, Z. H. Hakami, N. J. Prottasha, andM. Kowsher, "Computer-aided system for extending the performanceof diabetes analysis and prediction," in2021 International Conferenceon Software Engineering & Computer Systems and 4th InternationalConference on Computational Science and Information Management(ICSECS-ICOCSIM). IEEE, 2021, pp. 465–470

[11] Heng Wang, Dan Oneata, Jakob Verbeek, Cordelia Schmid. "A robust and efficient video representation for action recognition." International Journal of Computer Vision, Springer Verlag, 2016, 119 (3), pp.219–238.

[12] M. A. Rahim, M. Rahman, M. A. Rahman, A. J. M. Muzahid, and S. F.Kamarulzaman, "A framework of iot-enabled vehicular noise intensity-monitoring system for smart city,"Advances in Robotics, Automationand Data Analytics: Selected Papers from ICITES 2020, vol. 1350, p.194, 2021.

[13] Karen Simonyan, Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos." Advances in Neural Information Processing Systems. 1.

[14] S. Sudhakaran and O. Lanz, "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, 2017, pp. 1–6.

[15] B. Zhang, L. Wang, Z. Wang, Y. Qiao and H. Wang, "Real-Time Action Recognition With Deeply Transferred Motion Vector CNNs," in IEEE Transactions on Image Processing, vol. 27, no. 5, pp. 2326-2339, May 2018, doi: 10.1109/TIP.2018.2791180.

[16] L. C. Kiew, A. J. M. Muzahid, and S. F. Kamarulzaman, "Vehicle routetracking system based on vehicle registration number recognition usingtemplate matching algorithm," in2021 International Conference on-Software Engineering Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), 2021, pp. 249–254

[17] C. Feichtenhofer, A. Pinz and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1933-1941, doi: 10.1109/CVPR.2016.213.

[18] L. Wang, Y. Xiong, Z. Wang, Q. Zhe, L. Dahua, X. Tang, L.V. Xiaoou Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition." 9912. 10.1007/978-3-319-46484-8_2.

[19] Z. Lan, Y. Zhu, A. G. Hauptmann and S. Newsam, "Deep Local Video Feature for Action Recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1219-1225, doi: 10.1109/CVPRW.2017.161.

[20] B. Zhou, A. Andonian, A. Torralba, "Temporal Relational Reasoning in Videos." arXiv 2017, arXiv:1711.08496.

[21] J. Carreira, A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset." pp. 4724–4733, 10.1109/CVPR.2017.502.

[22] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga and G. Toderici, "Beyond short snippets: Deep networks for video classification," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4694-4702, doi: 10.1109/CVPR.2015.7299101.

[23] Jun Liu, Amir Shahroudy, Dong Xu, Gang Wang, "Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition." vol. 9907, 10.1007/978-3-319-46487-9_50.

[24] M. Kowsher, A. Tahabilder, and S. A. Murad, "Impact-learning: a robustmachine learning algorithm," inProceedings of the 8th internationalconference on computer and communications management, 2020, pp.9–13.

[25] A. J. M. Muzahid, S. F. Kamarulzaman, and M. A. Rahman, "Comparison of ppo and sac algorithms towards decision making strategies forcollision avoidance among multiple autonomous vehicles," in2021 In-ternational Conference on Software Engineering Computer Systems and4th International Conference on Computational Science and InformationManagement (ICSECS-ICOCSIM), 2021, pp. 200–205

[26] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In ICLR, 2015.

[27] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[28] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800-1807, doi: 10.1109/CVPR.2017.195.

[29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang,T. Weyand, M. Andreetto, H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017.

[30] C. Szegedy, S.Ioffe, V. Vanhoucke, A. A Alexander, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," AAAI Conference on Artificial Intelligence, 2016.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.

[32] D. Kingma,J. Ba, "Adam: A Method for Stochastic Optimization. International Conference on Learning Representations,",2014.

[33] A. Karim, M. A. Islam, P. Mishra, A. J. M. Muzahid, A. Yousuf,M. M. R. Khan, and C. K. M. Faizal, "Yeast and bacteria co-culture-based lipid production through bioremediation of palm oil mill effluent:a statistical optimization,"Biomass Conversion and Biorefinery, pp. 1–12, 2021

[34] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15). JMLR.org, 448–456, 2015.

[35] A. L. Maas, A. Y. Hannun, Andrew Y. Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models,".

[36] X. Glorot, Y. Bengio,"Understanding the difficulty of training deep feedforward neural networks," Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR 9:249-256, 2010.