Crime Detection in Surveillance Videos

Ashok Kumar J M, Arun Kumar C, Abishek B R, Thirumagal E

Abstract: — ***During the most recent couple of decades, surveillance cameras have been introduced in numerous areas. Examination of the data caught utilizing these cameras can assume powerful jobs in web based observing different occasion expectation and objective driven applications including inconsistencies and interruption identification. Wrongdoing has raised in our everyday lives, observation recordings are utilized to catch an assortment of true irregularities. Observing consequently a wide basic open zone is a test to be tended to. We can abuse ongoing PC vision calculations so as to supplant human work. The video observation framework is two-dimensional spatial data over a third measurement, that recognizes and predicts strange practices expecting to accomplish a shrewd reconnaissance idea. In this paper, we audit various methodologies used to learn inconsistencies by abusing both ordinary and atypical recordings. To abstain from clarifying the peculiar fragments or clasps in preparing recordings, which is very tedious, the learning calculation adapts irregularity through the different examples of positioning structures by utilizing the feebly marked preparing recordings.***

Index Terms: anomaly detection, surveillance systems, computer vision, feature extraction, object detection, object tracking, C3D, CNN, deep learning.

I INTRODUCTION

In computer vision, the analysis of images and videos was extremely difficult and the statistical models used worked only for a handful of tasks. The introduction of Deep Learning approaches has revolutionized the field. Given enough labelled data it is virtually possible to solve any vision-related problem. In 2012 when [1] Krizhevsky et al., were able to win the ImageNet LSVRC competition, everyone was amazed to see the power of convolutional neural networks. The novel idea of convolutions was introduced by [2] LeCun et al., way back in 1998, which is currently the basic unit used to solve all kinds of computer vision tasks using deep learning methods.

Anomaly indicates events which are not usual, not regular, not expected and not predictable and so it is different from existing designs [3]. So the task of detecting an anomaly involves recognition of different situations from the video clip, where the anomaly activity takes place only during a small time period.

Revised Manuscript Received on April 24, 2019
Ashok Kumar J M, REVA University, India.
Arun Kumar C, REVA University, India.
Abishek B R, REVA University, India.
Thirumagal E REVA University, India.

In spite of the great success of deep learning models in tasks like image recognition, image classification, and object localization, these 2D

convolutional neural network models don't perform well with video analysis. Some of the reasons being, they are not able to capture the context that is continuous from one frame to another, they are computationally very expensive and the lack of architecture to capture temporal information. So we will be using a 3D convolutional neural network for our experiments.

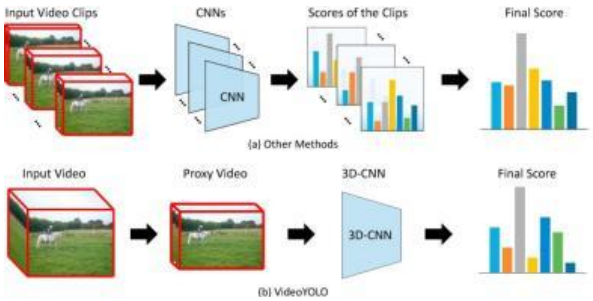


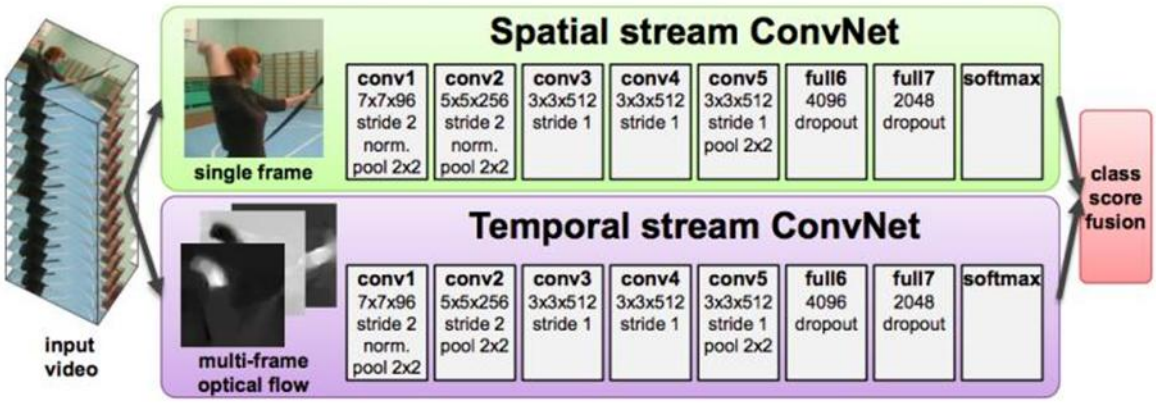
Fig: 1 2D CNN vs 3D CNN [4]

The number of surveillance cameras installed at streets, shopping malls, banks, parks, etc. is growing day by day in the interest of public safety. The data generated by these cameras is in petabytes, but there is not enough human labor to analyze these videos as the human to camera ratio is very low in most places. Another problem is that human involvement introduces problems of selection bias and indigeneity [5]. An automated intelligent anomaly detection system will help in solving some of these problems.

Crime is a very broad word, as a lot of things count as crime and teaching what crime is to your model is going to be difficult. One approach is to develop task-specific models like violence detection [6,7] and traffic crime detection [8,9,10], but these models are task specific and would not work as a general crime detection system that is needed in the real world. Another approach is to train your model on multiple crimes so that it is able to recognize crime as a whole and not the particular crime that is trained on.

Real world crimes are unique and different each time, so it is very important for the system to be trained on a diverse dataset that helps it to generalize what crime is. This kind of learning can be achieved using either.

Real world crimes are unique and different each time, so it is very important for the system to be trained on a diverse dataset that helps it to generalize what crime is. This kind of learning can be achieved using either semi-supervised [12] or unsupervised learning [11]. For unsupervised learning, we



would need a lot of diverse data and it is very difficult to find a dataset that covers all kinds of crime, hence our model is based on a semi-supervised learning approach. The dataset used in the model is

Fig: 2 Two stream architecture [23]

UFC-Anomaly Detection Dataset, as introduced in [13]. The model uses two-stage approach, wherein the first stage features from the video are extracted and these features are then given as input to the second model which is a simple fully connected neural network, the workings of which will be explained in the later section.

This paper constitutes of the sections as follows: Section **II** discuss the related works. Section **III** elaborates on the existing methodologies and architecture. Section **IV** discusses the experiment and results, preprocessing, cross-validation, data augmentation, the parameter of training, results and quantitative analysis. Section **V** discuss the conclusions and the enhancements that can be made in future.

II. RELATED WORKS

In computer vision, video processing is a classic problem which has profound applications in various fields such as crime detection, action recognition, medical images like MRI, etc. These have been significant advancements in image processing using deep learning methods but the same cannot be said for video analysis. Models like AlexNet[14], ResNet[15], VGG[16], DenseNet[17] and GoogleLenNet[18], have changed the field of image processing for the good. But the same can't be said for video processing. There have been some interesting papers published recently which we will be discussing below.

There are a few reasons that make video processing more difficult than image processing. Computational expense being one noteworthy reason, a basic 2D convolutional neural system that does classification on 101 classes has quite recently around 5 million parameters while a similar architecture when inflated to a 3D structure results in roughly 33 million parameters. It takes around 72 to 96 hours to prepare a 3D convolutional neural system on UCF101[19] dataset and around two months on Sports-1M [20] dataset, this makes architecture look troublesome and overfitting likely [21]. Another problem being capturing the spatiotemporal context over a long time period. There was also a lack of video processing specific architectures as well as standard benchmarks that made it difficult, but of late these problems have been solved. There are two main approaches to video processing. Single stream network [22] introduced in 2014, joined the temporal information from multiple continuous frames in different ways. Although a novel approach, it could not capture the temporal information that well. Two Stream Networks [23] as shown in Fig: 2, was built on the failures of single stream network. It had a separate convolutional neural network stream to capture the time related information. These are two backbones on which all the future research was built on.

Long term Recurrent Convolutional Networks for Visual Recognition and Description [24] utilized Long short-term memory [25] units that are accustomed to process temporal information, to train the model on independent features to test if they capture time related information in videos.

In Learning Spatiotemporal Features with 3D Convolutional Networks [26], they built upon the works of Karpathy et al [22] to improve results. The primary thought behind this paper is that to train a huge 3D convolutional neural system on tremendous datasets like Sports-1M [20] and utilize these models as highlight extractors for different datasets. They found that a straightforward neural system or a direct classifier like Support Vector Machine would give preferred outcomes over a solitary model. They found that that the model concentrated on spatial appearance in the initial couple of frames and followed the movement in the frames that are upcoming.

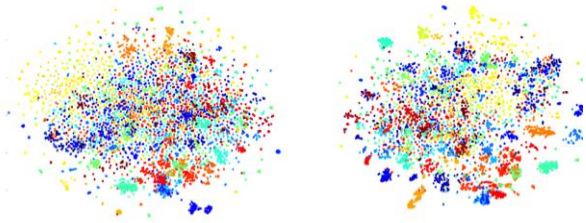


Fig: 3 On the left Clustering based on ImageNet, on the right clustering based on C3D [26]

As we see in Fig: 3 the C3D model is able to cluster different classes better than models trained on ImageNet. Due to the robustness of the model, good documentation and availability of good implementation, we have used C3D as our first stage model to extract features from the videos. There have been significant improvements in video processing after C3D, like Describing Videos by Exploiting Temporal Structure [27], Convolutional Two-Stream Network Fusion for Video Action Recognition [28], Temporal Segment Networks: Towards Good Practices for Deep Action Recognition [29], Action VLAD: Learning

spatiotemporal total for activity arrangement [30], Hidden Two-Stream Convolutional Networks for Action Recognition [31], Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset [32], Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification [33], A Closer Look at Spatiotemporal Convolutions for Action Recognition [34]and a lot more such models have been released that use different approaches to solve the different problems related video analysis. Crime detection is one of the major research fields and a lot of people are working on it [13, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46]. They have focused on different tasks using different approaches. There are also focused work in violent and aggressive crime detection [47,48,49,6], which are really interesting. There is some focused work on traffic crime detection [8,9,10] as well that had interesting approaches. As we see in [13] the ranking based approach used is very effectively been used for crime detection and the same is also demonstrated by [46, 35].

III. METHODOLOGY

The classic solution would be to use a 2D convolutional neural network and process one frame at a time and then combine the output of those frames to get the anomaly result, but that doesn't work as well as a 3D convolutional neural network-based model. We first reduce the dimensions of the video to 240×320 dimension pixels and reduces the rate of frames to 30 fps. Then we divide each video into multiple smaller clips, with each clip containing 32 frames. They are stored in two folders, one containing positive clips and the second containing negative clips.

The advent of transfer learning in deep learning has given a boost to researchers as it allows them to transfer the knowledge of a model trained on one dataset to another model. Here C3D is trained on the humongous Sports-1M dataset and the same model is successfully able to extract features from crime videos as well. After the feature extraction, we store these features in a text file. These features are to be fed into 3-layer neural network, with the first layer having 4096 neurons, the second layer having 512 neurons and the last layer having 32 neurons. Dropout of 0.6 is used after each layer to add normalization. The activation function used is ReLU for the first 3 layers and the final layer uses Sigmoid activation function. The loss function used is MIL [13]. The output is a number, a score of an anomaly being present in the clip processed.

IV. EXPERIMENT AND RESULTS

The proposed model has been run on Windows version 10 with 16 GB of RAM base on Keras deep learning framework using python 3.7. Other software requirements include Theano, Matplotlib, Numpy. We have utilized the FC-Anomaly Detection Dataset that contains 95 GB of videos in total. The dataset and the C3D tool was installed on Google Cloud Platform, as the pre-processing needs, a lot of computational power and cloud services like GCP help us fulfil this requirement. First, all the videos are downloaded onto the cloud instance and then C3D is installed and run on to generate features on selected videos. Once the features are extracted into a text file they are downloaded and used on the second model to train and test them.

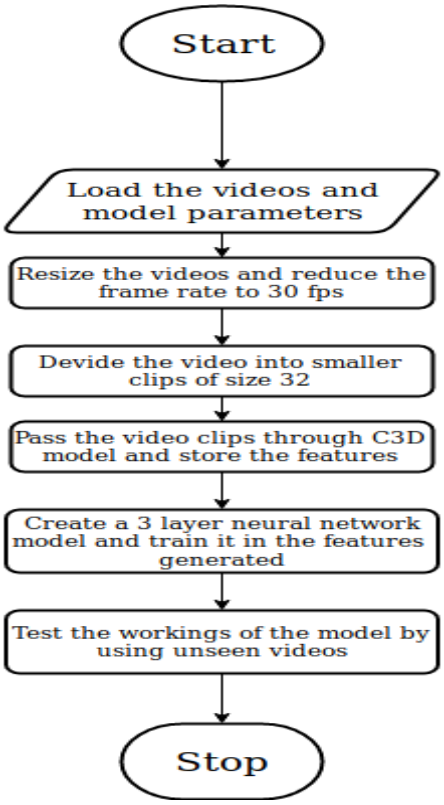


Fig: 4 Flowchart of the used approach in this paper

A. Pre-Processing

This step is mandatory to expand the amount of training data and helps achieve better computational efficiency. As the name goes, the data before being fed into the network would undergo processing like color normalization, resizing and clipping.

.

B. The Parameter of training

To avoid the problem of overfitting in deep learning we need to make sure that a good amount of qualitative data is used while training. The use of dropout also helps in making sure that your model is not overfitting. The parameters of the C3D models are fixed in this implementation, but there is a lot of scopes expected in improving the model performance by training both the C3D model along with a neural network model and passing the error all the way through till the first layer of C3D model.

C. Results and Quantitative Analysis

The Performance of the model is evaluated using the AUC - ROC Curve. ROC does the probability measurement and AUC does the separability measurement. It predicts the capability of the model in distinguishing between classes

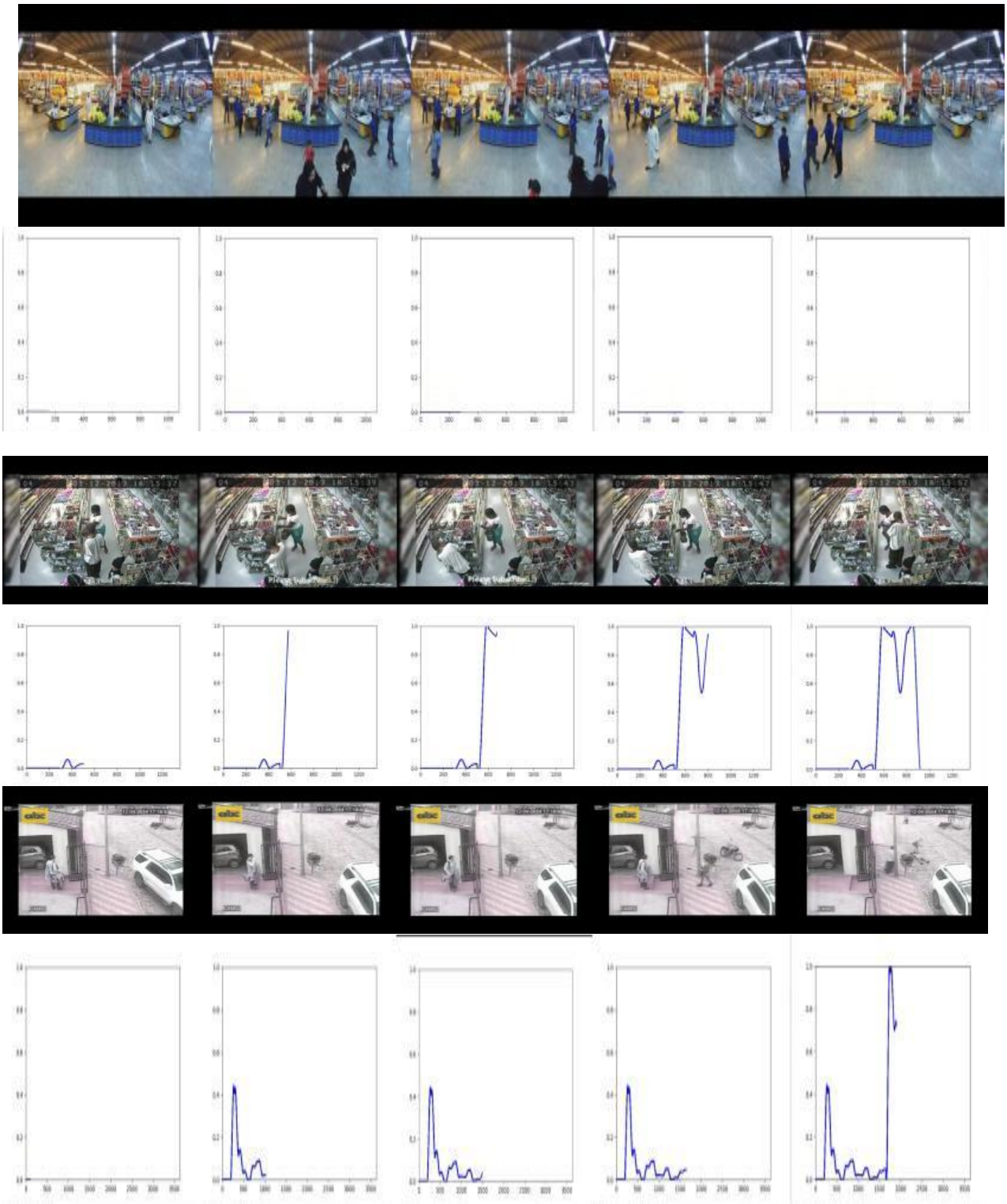


Fig: 5 Top to bottom, screenshots of normal activity, screenshots of shoplifting, screenshots of shooting

We were able to get 0.72 scores on this matric. The sample results are shown in Fig: 4. As we can see on the top, in normal activity the model does not suspect any crime. Next, we see that a lady is shoplifting and the model is clearly able to detect that and the anomaly score goes high. In the last row, we see that in the last two frames a person is shot at and the criminals running away. The model is successfully able to detect when the shooting occurs, but it has also raised a false alarm at the starting of the video.

Such two-stage video processing models have found significant applications not only in video analysis but also in the medical field to process MRI images, Motion tracking, etc. and is an interesting field of research.

V. CONCLUSION & FUTURE WORK

The paper demonstrates that a two-stage approach for Crime Detection in Surveillance Videos. The C3D model is able to acquire the features and structures from the videos successfully and a simple neural network on top is able to get good results. In future work, it is planned to implement and investigate different deep learning models that could be used as feature extractors. Additionally, it is been planned to investigate and possibly implement deep learning end-to-end models that take video as input and give anomaly score as output.

Given the biggest obstacle in the video, an analysis is a computational cost, the future work must focus on speeding up the process and reducing the computational time, enabling research to try different models faster and cheaper. Some novel approaches like capsule nets [50] should also be tested with video processing. There should also be research in collecting more diverse data, that covers different kinds of crimes and collecting geographically diverse data so that the model is able to generalize better and have minimal or no bias.

VI. ACKNOWLEDGMENT

The authors would like to thank REVA University Bengaluru, India for providing all the support for the development of the cyber physical system and school of Computing and Information technology for the support extended.

VII. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", NIPS Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 Pages 1097-1105, 2012.

[2] Yann LeCun, Leon Bottou, Yoshua Bengio, Patrick Haner, "Gradient-Based Learning Applied to Document Recognition", Proceedings of the IEEE (Volume: 86, Issue: 11, Nov 1998).

[3] Dinesh Kumar Saini, Dikshika Ahir, Amit Ganatra, "Techniques and Challenges in Building Intelligent Systems: Anomaly Detection in Camera Surveillance", *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems*, 2016.

[4] Longlong Jing, Xiaodong Yang, Yingli Tian, "Video you only look once: Overall temporal convolutions for action recognition", Journal of Visual Communication and Image Representation, Volume 52, April 2018, Pages 58-65.

[5] Mikael Priks, "The Effects of Surveillance Cameras on Crime: Evidence from the Stockholm Subway". The Economic Journal, Volume 125, Issue 588, November 2015, Pages F289–F305.

[6] S. Mohammadi, A. Perina, H. Kiani, M. Vittorio, "Angry crowds: Detecting violent events in videos", ECCV, 2016.

[7] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, Anderson Rocha, "Temporal Robust Features for Violence Detection", IEEE Winter Conference on Applications of Computer Vision (WACV), 2017.

[8] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi, "Traffic monitoring and accident detection at intersections", IEEE Transactions on Intelligent Transportation Systems, Volume: 1, Issue: 2, Page(s):108–118, Jun 2000.

[9] Xiaohui Huang, Pan He, Anand Rangarajan, Sanjay Ranka, "Intelligent Intersection: Two-Stream Convolutional Networks for Real-time Near Accident Detection in Traffic Video", ACM Transactions on Spatial Algorithms and Systems (TSAS), 2019.

[10] Yu Yao, Mingze Xu, Yuchen Wang, David J. Crandall, Ella M. Atkins, "Unsupervised Traffic Accident Detection in First-Person Videos", arXiv preprint arXiv:1903.00618, 2019.

[11] Alec Radford, Luke Metz, Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", arXiv preprint arXiv:1511.06434, 2016.

[12] Rie Johnson and Tong Zhang, "Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding", Advances in Neural Information Processing Systems 28 (NIPS 2015).

[13] Waqas Sultani, Chen Chen, Mubarak Shah, "Real-world Anomaly Detection in Surveillance Videos", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[15] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, Andrew Y N, "Reading Digits in Natural Images with Unsupervised Feature Learning", NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

[16] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", International Conference on Learning Representations (ICLR), 2015.

[17] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, "Densely Connected Convolutional Networks", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, "Going Deeper with Convolutions", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[19] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "UCF101: A Dataset of 101 Human Action Classes from Videos in The Wild.", CRCV-TR-12-01, November 2012.

[20] Andrej Karpathy and George Toderici and Sanketh Shetty and Thomas Leung and Rahul Sukthankar and Li Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[21] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, Manohar Paluri, "ConvNet Architecture Search for Spatiotemporal Feature Learning", arXiv preprint arXiv:1708.05038, 2017.

[22] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Li Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks", IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2014.

[23] Karen Simonyan and Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", Advances in Neural Information Processing Systems 27 (NIPS 2014).

[24] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 39, Issue: 4, Page(s): 677 - 691, 2017.

[25] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory", Neural Computation, Volume 9, Issue 8, November 15, 1997, p.1735-1780.

[26] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks", IEEE International Conference on Computer Vision (ICCV) Pages 4489-4497, 2015.

[27] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, Aaron Courville, "Describing Videos by Exploiting Temporal Structure", IEEE International Conference on Computer Vision (ICCV), 2015.

[28] Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[29] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, Luc Van Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition", European Conference on Computer Vision (ECCV), 2016.

[30] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, Bryan Russell, "ActionVLAD: Learning spatiotemporal aggregation for action classification", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[31] Yi Zhu, Zhenzhong Lan, Shawn Newsam, Alexander G. Hauptmann, "Hidden Two-Stream Convolutional Networks for Action Recognition", Asian Conference on Computer Vision (ACCV), 2018.

[32] Joao Carreira and Andrew Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", arXiv preprint arXiv:1705.07750, 2017.

[33] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, Luc Van Gool, "Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification", arXiv preprint arXiv:1711.08200, 2017

[34] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, Manohar Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[35] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, Nicu Sebe, "Learning Deep Representations of Appearance and Motion for Anomalous Event Detection", British Machine Vision Conference (BMVC), 2015

[36] Shandong Wu, Brian E. Moore, Mubarak Shah, "Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[37] Arslan Basharat, Alexei Gritai, Mubarak Shah, "Learning Object Motion Patterns for Anomaly Detection and Improved Object Detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[38] Xinyi Cui, Qingshan Liu, Mingchen Gao, Dimitris N. Metaxas, "Abnormal detection using interaction energy potentials", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[39] Borislav Anti´c and Bj¨orn Ommer , "Video Parsing for Abnormality Detection", International Conference on Computer Vision(ICCV), 2011.

[40] Timothy Hospedales, Shaogang Gong, Tao Xiang, "A Markov Clustering Topic Model for Mining Behaviour in Video", International Conference on Computer Vision(ICCV), 2009.

[41] Yingying Zhu, Nandita M. Nayak, Amit K. Roy-Chowdhury, "Context-Aware Activity Recognition and Anomaly Detection in Video", IEEE Journal of Selected Topics in Signal Processing, 2013.

[42] Weixin Li, Weixin Li, Weixin Li, "Anomaly Detection and Localization in Crowded Scenes", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 36 Issue 1, January 2014

[43] Louis Kratz, Ko Nishino, "Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[44] Cewu Lu, Jianping Shi, Jiaya Jia, "Abnormal Event Detection at 150 FPS in MATLAB", International Conference on Computer Vision(ICCV), 2013.

[45] Bin Zhao, Li Fei-Fei, Eric P. Xing, "Online Detection of Unusual Events in Videos via Dynamic Sparse Coding", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[46] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, Larry S. Davis, "Learning Temporal Regularity in Video Sequences", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[47] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, Yi Liu, "Violence detection using Oriented VIolent Flows", Image and Vision Computing, 2016.

[48] J.F.P. Kooija, M.C. Liem, J.D. Krijnders, T.C. Andringa, D.M.Gavrila, "Multi-modal human aggression detection", Computer Vision and Image Understanding, 2016.

[49] Ankur Datta, Mubarak Shah, Niels Da Vitoria Lobo, "Person-on-Person Violence Detection in Video Data", International Conference on Pattern Recognition (ICPR), 2002.

[50] Sara Sabour, Nicholas Frosst, Geoffrey E Hinton, "Dynamic Routing Between Capsules", Advances in Neural Information Processing Systems 31 (NIPS 2017).

AUTHORS PROFILE

Ashok Kumar J M is final year student in the department of Computer Science and Engineering in REVA University, Bengaluru, India

Abishiek B R is final year student in the department of Computer Science and Engineering in REVA University, Bengaluru, India

ARUN KUMAR C is final year student in the department of Computer Science and Engineering in REVA University, Bengaluru, India

Mrs.Thirumagal E is an ACM and CSI member and currently working as assistant professor in the department of Computer Science and Engineering in REVA University, Bengaluru, India