IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Anomalous Human Action Recognition With Deep Learning Technique

Nazneen, Prof.Shruthi S A

M.Tech Student CSE Department, Assistant Professor CSE Department.

Department of Computer Science and Engineering

Faculty of Engineering & Technology (Exclusively For Women), Sharnbasva University, Kalaburagi, Karnataka, India

Abstract— One of the cornerstones of computer vision research is human activity recognition. Robotics relies heavily on HAR accuracy. Using a YOLOv7-based model for human action recognition is the focus of this book chapter. Results from action recognition using YOLOv7 and CNN+LSTM were evaluated to assess efficacy of model. Additionally, we provide a compact human action dataset that is well-suited for training YOLO models. The pictures in this collection came from the KTH, Weizmann, and MSR archives. We utilise this data set to confirm the experimental findings in this study. YOLOv7 model was shown to be effective for HAR in final testing findings. Anomaly detection, which may spot crimes including robbery, assault, fighting, and arrest, is carried out in this research. Anomaly events like Robbery depict robbers illegally obtaining money by force or fear of force, and Abuse shows films of horrible, aggressive, or abusive behaviour against women, children, the elderly, and animals. No gunfire is shown in these films. There are videos called "Shooting" that show someone shooting another person, "Fighting" that show two or more people fighting, and "Arrest" that show police taking people into custody.

***Keywords*—YOLOv7, CNN, LSTM, Abuse, Robbery, Shooting**

I. INTRODUCTION

One of most important subfields in broader field of computer vision & AI is study of abnormal human action detection, which seeks to detect out-of-the-ordinary behaviours in video clips. Identifying aberrant or unexpected behaviours is crucial in many applications, such as autonomous driving, & surveillance, where it may greatly affect operational effectiveness and security. Advent of deep learning has been a game-changer in this area, as it has allowed for the development of more effective anomaly detection systems by giving new ways to evaluate and understand complicated patterns in video data.

Recognising unusual human behaviour is difficult because people and their environments are inherently unpredictable. Inadequacy of traditional methods to adequately represent contextual information and temporal dynamics meant that they often failed to meet these difficulties. By allowing extraction of rich spatial and temporal data from video sequences, deep learning—specifically via CNNs and RNNs—has overcome these restrictions. Detailed spatial patterns linked to human behaviours may be captured by CNNs because of their proficiency in learning hierarchical representations of visual input. To the contrary, RNNs—and in particular, LSTM networks—are great at representing sequential information and temporal dependencies, that are fundamental for comprehending how actions change over time.

Complex designs like 3D CNNs have recently been proposed, allowing for the processing of spatiotemporal data by expanding the classic 2D convolutional models. 3D CNNs capture spatial and temporal dynamics more efficiently by evaluating several frames concurrently, giving a more complete picture of the activity. Capacity of Transformer-based models to manage contextual linkages and long-range dependencies in video sequences has also contributed to their rise to popularity. By using processes such as self-attention, these models are able to enhance their ability to identify tiny abnormalities by constantly adjusting their attention to important areas and focusing on various sections of sequence.

The use of attention processes has improved anomaly detection systems even further by training the algorithm to focus on high-risk regions of the video. Detection accuracy and false positive rate are both enhanced by this focused strategy. Additionally, methods like transfer learning and self-supervised learning are being investigated as potential ways to enhance the performance of models, particularly in cases when there is a lack of labelled data.

It is common practice to use massive datasets that include a wide variety of normal and abnormal human activities when training deep learning models to detect anomalous actions. The generalizability and robustness of the models are improved by the use of synthetic data creation and data augmentation. With an emphasis on obtaining high accuracy in identifying anomalies while minimising false detections, these models are assessed using metrics.

Recognising abnormal human actions has many practical uses. When it comes to security and surveillance, being able to spot suspicious activity in real-time is crucial for thwarting any dangers and improving knowledge of the current situation. By keeping tabs on patients, healthcare providers may see any unusual behaviour that might be a sign of an emergency. Anomaly detection helps autonomous cars improve their dependability and safety by allowing them to react to unexpected movements made by people or other vehicles.

However, there are still a number of obstacles to overcome. Existing systems still have a hard time dealing with real-time processing, environmental changes, and individual behaviour variations. Better data gathering methods, more efficient models, and new anomaly detection approaches are the focus of ongoing research to overcome these issues.

Finally, the area of anomalous human action identification using deep learning approaches is one that is both exciting and quickly changing, with big consequences for practical uses. More intelligent and responsive technology in many fields are on the horizon, thanks to researchers' efforts to improve anomaly detection systems using cutting-edge neural network designs and novel approaches.

II. Related Work

Yan, M., Meng, et.al [6], 2020, The unusual look and motion patterns, known as spatiotemporal irregularities, may be hard to spot in recordings since they aren't always clearly visible and don't happen very often. To solve this challenge, we train our system to recognise typical patterns in films and label outliers as such. In order to do this, we present a 3D-FCAE that can be trained end-to-end and can identify spatiotemporal and temporal anomalies in films with little training data. Consequently, frames with large reconstruction errors indicate temporal inconsistencies, whereas hazy, poorly reconstructed portions indicate uneven spatiotemporal patterns. With the help of the 3D fully convolutional autoencoder and the investigated effective architecture, our method is able to precisely pinpoint spatiotemporal and temporal imperfections. We test suggested autoencoder on benchmark video datasets with little supervision in order to identify anomalous patterns. Our approach's efficacy is shown by comparisons with state-of-the-art methods. In addition, the learnt autoencoder performs well when applied to different datasets.

Wang, T et.al.[8], 2020, When it comes to everyday personal protection, security monitoring of public scenes is quite crucial. This worry is driving the rise of anomaly detection as a top priority in the fields of computer vision and video processing. Our goal in this research is to provide a novel method that can identify visual abnormalities. First, our technique separates the issue into its component parts. Then, we use an AutoEncoder-based network termed CDA to extract feature descriptors. A new descriptor that captures multi-frame optical flow data represents movement information. Prior to training the CDA network, feature descriptor of normal samples is inputted. As a last step in testing process, CDA reconstruction error is used to differentiate aberrant samples. Using several video surveillance datasets, we verify suggested strategy.

[9] Hu, Z., et.al. Parallel [9], 2020, Growing use of intelligent surveillance to bolster public safety has piqued interest of the computer vision research community in identification and localisation of anomalies in densely populated environments. We provide new model of parallel spatial-temporal CNN to identify and pinpoint suspicious activity in surveillance footage. There are two primary components to our method. First, we provide a new spatial-temporal cuboid of interest identification approach using optical flow algorithm and cells of varying sizes, taking into account usual camera location and a huge amount of background information. Next, we explain same behaviour in various temporal-lengths using parallel 3D CNN. That is first stage in capturing cuboids' behaviour data, and it also reduces data that isn't relevant to main behaviour. We outperformed state-of-the-art approaches in evaluations conducted on benchmark datasets.

For the purpose of human action recognition, more refined deep networks were developed (Gowdra, et al. 2021). Following YOLOv3, more sophisticated approaches were created to address the issue of human action recognition (Lu et al., 2020). The findings are obtained by this novel approach using the network architecture YOLOv4 + LSTM. The network's accuracy reached 97.87% after numerous trials. For this reason, the recognition technique takes into account both geographical and temporal data. They went a step farther by empirically validating an attention-based Selective Kernel Network (SKNet) model. Results improved with this model's application. Different approaches exist for identifying human actions.

Wang et al., for instance, used human gait identification based on multichannel convolutional neural networks (Wang, Zhang, & Yan, 2020) and frame-by-frame detection of human gait energy pictures (Wang & Yan, 2020; Liu & Yan, 2020). In the end, both approaches produced promising experimental outcomes. Nevertheless, YOLO-based human action recognition approaches are still in their infancy. The majority are using different deep learning methods, such as CNN.

To classify the actions and behaviours of many human subjects, Nguyen & Bui (2023)[21] presented a new method that combines deep learning and machine learning approaches. The proposed approach comprises a three-stage identification process that starts with object detection using the YOLOv5 model, continues with skeletal visualisation using a media pipeline, and concludes with action identification using an LSTM network. The results of the YOLOv5 target detection experiments show that the loss accuracy stays below 5% throughout validation and training. For every one of the case studies that were considered, the YOLOv5 model's mean accuracy (mAP) was more than 99%.

III. METHODOLOGY

In order to identify human actions, this technique uses the YOLOv7 architecture.Manufacturing may be made more efficient and precise with application of YOLOv7 algorithms, which could identify and follow things as they travel down production line. Furthermore, object detection is utilised to identify manufacturing defects and ensure product or component quality.

Alogrithm Steps:

Steps for object Detection using YOLO v7:

- Shape (m, 416, 416, 3), represented as a batch of photos, serves as the input.
- Convolutional neural networks (CNNs) are fed this picture via YOLO v7.
- In order to get an output volume of (19, 19, 425), the final two dimensions of the previous output are flattened:
- There is a 19 × 19 grid, and 425 numbers are returned for each cell.
- There are 5 anchor boxes in each grid, thus 425 is equal to 5 times 85.
- Five is the set of possible values (pc, bx, by, bh, bw), and eighty is the required number of classes (85 = 5 + 80).
- The final product includes the identified classes and a list of bounding boxes. Numbers pc, bx, by, bh, bw, and c represent each of the six bounding boxes. Each boundary box is represented by 85 integers when 'c' is stretched into an 80-dimensional vector.
- Lastly, in order to prevent boxes from being selected that overlap, IoU (Intersection over Union) and Non-Max Suppression are used.

IV. SYSTEM ARCHITECTURE

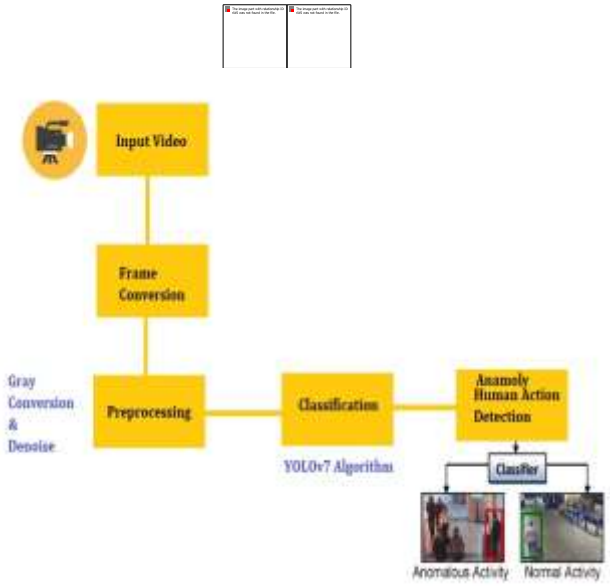


Figure 1: System Architecture

Before applying YOLO, system under aforementioned architecture transforms video to frames and does preparatory operations like grayscale conversion and denoising.

1. 1. Model Architecture

YOLOv7 improves upon and streamlines its predecessors in terms of architecture. It comprises parts such as:

- **Backbone Network**: A CNN which takes pictures as input and uses them to create a model. CSP networks are used by YOLOv7 to improve gradient flow and decrease computing complexity.
- **Neck**: Rich feature representations are generated by the neck by aggregating feature maps from various levels of the backbone. The PANet is used by YOLOv7 to accomplish this.
- **Head**: Estimates of objectness scores, class probabilities, and bounding boxes are made by the detection head.

2. Input Processing

Image Resizing: To make sure input size is what network is expecting, input picture is shrunk to a set dimension, such 640x640 pixels.

Normalization: Pixel values are normalized to a specific range.

3. Feature Extraction

Feature maps at various scales are extracted from the input picture by backbone network. Feature maps like this may capture everything from most minute details to the most generalised ideas.

4. Feature Aggregation

Feature maps from various scales are aggregated and fused in the neck. By enhancing data flow and creating feature pyramids, PANet aids in the detection of objects of varying sizes.

5. Prediction

Predictions are made by head network using aggregated feature maps. With the use of a one-stage detector, YOLOv7 is able to estimate the class probabilities, objectness scores, and bounding boxes of several anchor boxes in every grid cell.

6. Bounding Box Prediction

Anchor Boxes: Predefined anchor boxes (priors) are used to predict bounding boxes. The network adjusts these anchors to fit the objects in the image.

Bounding Box Regression: So that anchor boxes may be moved and resized, the network makes predictions about offsets.

Objectness Score: Network makes an educated guess as to likelihood that anchor box has an item.

Class Probabilities: Network makes predictions about bounding box object's class probabilities.

7. Non-Maximum Suppression (NMS)

Once YOLOv7 has predictions, it uses NMS to get rid of unnecessary bounding boxes. NMS retains top-scoring boxes while removing those having high intersection over union (IoU) with them..

8. Post-Processing

Last set of discoveries consists of remaining bounding boxes, their confidence ratings, & labels assigned to each class.

Size of bounding boxes is returned to initial picture size.

9. Loss Function

With a bespoke loss function which includes numerous components, YOLOv7 achieves its goal:

Localisation loss is the difference in accuracy among ground truth boxes & expected bounding boxes.

Objectness score prediction error is quantified as confidence loss.

Calculates inaccuracy in forecasting class probabilities; this is known as class probability loss.

10. Training Process

SGD and its derivatives are used to train YOLOv7 on large-scale datasets, such as COCO. To make model more resilient, data augmentation methods including random scaling, cropping, & flip are utilised.

11. Inference

Inference is technique by which trained YOLOv7 model uses picture processing to identify objects in real-time. If your application needs object recognition quickly and accurately,

YOLOv7 is way to go because of how fast and efficient it is.

V. IMPLEMENTATION

Figure 2: Menu

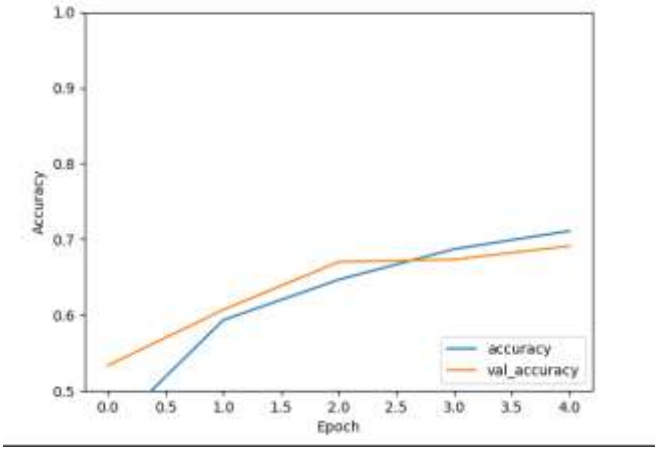
Capture video, convert frames, segment, extract features, and activate action recognition are all available in this menu module.

Fig-3: Read Image

This component takes video footage and pulls out the frame.

Fig-4: Human action recognition

Human activity is identified by this module.



Graph 1: Epoch vs Accuracy graph
Displays the graph of Epoch Vs Accuracy

VI. CONCLUSION

We proposed Yolo as a means of detecting human behaviour. First evaluation of HAR with YOLOv7 yields positive findings. First evidence whether YOLOv7 may be utilised for HAR is shown in this paper. With goal of making it compatible with YOLOv7 model, a new action recognition dataset is constructed utilising existing datasets. Our current findings demonstrate that six preset behaviours may be accurately recognised & categorised.

Future scope

In addition, the model will have pre-judgment conditions added to it. We derived this rationale from our experiments. If identified picture has more than one tag. For instance, during a video's action, numerous tags may exist at once or shift between them. When you add this condition, action will receive recognition only if all of requirements are satisfied. In order to further assess YOLOv7-based human action recognition algorithms, we will keep looking for appropriate datasets.

REFERENCES

[1] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 fps in Matlab,” in Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, 2013, pp. 2720–2727.

[2] Wang, T., Chen, J., & Snoussi, H. Online detection of abnormal events in video streams. Journal of Electrical and Computer Engineering, 2013.

[3] Lee, S. C., & Nevatia, R, Hierarchical abnormal event detection by real time and semi-real time multi-tasking video surveillance system. Machine Vision and Applications, 25(1), 133-143, 2014.

[4] Gu, X., Cui, J., & Zhu, Q, Abnormal crowd behaviour detection by using the particle entropy. Optik-International Journal for Light and Electron Optics, 125(14), 3428-3433, 2014.

[5] Varol, G., & Salah, A. A, Efficient large-scale action recognition in videos using extreme learning machines. Expert Systems with Applications, 42(21), 8274-8282, 2015.

[6] Yan, M., Meng, J., Zhou, C., Tu, Z., Tan, Y., Yuan, J. Detecting Spatiotemporal Irregularities in Videos via a 3d Convolutional Autoencoder. Journal of Visual Communication and Image Representation, 2020, 67, 47-59. https://doi.org/10.1016/j.jvcir.2019.102747

[7] Zhu, X., Liu, J., Wang, J., Li, C., Lu, H. Sparse Representation for Robust Abnormality Detection in Crowded Scenes. Pattern Recognition, 2014, 47(5), 1791-1799. https://doi.org/10.1016/j.patcog.2013.11.018

[8] Wang, T., Qiao, M., Zhu, A., Shan, G., Snoussi, H. Abnormalevent Detection via the Analysis of MultiFrame Optical Flow Information. Frontiers of Computer Science in China, 2020, 14(2), 304-313. https:// doi.org/10.1007/s11704-018-7407-3

[9] Liu, W., Luo, W., Lian, D., Gao, S. Future Frame Prediction for Anomaly Detection - A New Baseline. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. https://doi.org/10.1109/ CVPR.2018.00684

[10] Hu, Z., Zhang, L., Li, S., Sun, D. Parallel Spatial-Temporal Convolutional Neural Networks for Anomaly Detection and Location in Crowded Scenes. Journal of Visual Communication and Image Representation, 2020, 67, 765-771. https://doi.org/10.1016/j.jvcir.2020.102765

[11] Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D. Robust Real-Time Unusual Event Detection Using Multiple Fixed-Location Monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(3), 555-560. https://doi.org/10.1109/TPAMI.2007.70825

[12] Afiq, A. A., Zakariya, M. A., Saad, M. N. M., Nurfarzana, A. A., Khir, M. H. M., Fadzil, A. F., Jale, A., Gunawan, W., Izuddin, Z. A. A., Faizari, M. A Review on Classifying Abnormal Behavior in Crowd Scene. Journal of Visual Communication and Image Representation, 2019, 58, 285-303. https://doi.org/10.1016/j.jvcir.2018.11.035

[13] Ben, M., A., Zagrouba, E. Abnormal Behavior Recognition for Intelligent Video Surveillance Systems: A Review. Expert Systems with Applications, 2018, 91(1), 480-491. https://doi.org/10.1016/j.eswa.2017.09.029

[14] Almazroey, A. A., Jarraya, S. K. Abnormal Events and Behavior Detection in Crowd Scenes Based on Deep Learning and Neighborhood Component Analysis Feature Selection. In: Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), 2020, 258-267. https://doi. org/10.1007/978-3-030-44289-7_25

[15] Calderara, S., Heinemann, U., Prati, A., Cucchiara, R., Tishby, N. Detecting Anomalies in People's Trajectories Using Spectral Graph Analysis. Computer Vision and Image Understanding, 2011, 115(8), 1099-1111. https://doi.org/10.1016/j.cviu.2011.03.003

[16] Chang, H., Wang, T., Li, A., Fang, H. Local Hyperspectral Anomaly Detection Method Based on Low-Rank and Sparse Matrix Decomposition. Journal of Applied Remote Sensing, 2019, 13(02), 1-8. https://doi. org/10.1117/1.JRS.13.026513

[17] Elbayoudi, A., Lotfi, A., Langensiepen, C. The Human Behavior Indicator: A Measure of Behavioural Evolution. Expert Systems WITH Applications, 2019, 118, 493-505. https://doi.org/10.1016/j.eswa.2018.10.022

[18] Ergen, T., Kozat, S. S. Unsupervised Anomaly Detection with LSTM Neural Networks. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(8), 3127-3141. https://doi.org/10.1109/TNNLS.2019.2935975

[19] Febin, I. P., Jayasree, K., Joy, P. T. Violence Detection in Videos for an Intelligent Surveillance System Using Mobsift and Movement Filtering Algorithm. Pattern Analysis and Applications, 2020, 23, 611-623. https:// doi.org/10.1007/s10044-019-00821

[20] Dwivedi, N., Singh, D. K., Kushwaha, D. S. Orientation Invariant Skeleton Feature (oisf): A New Feature for Human Activity Recognition. Multimedia Tools and Applications, 2020, 79, 1-36. https://doi.org/10.1007/ s11042-020-08902-w

[21]Nguyen, A. T., & Bui, H. A. (2023, March). Multiple target activity recognition by combining YOLOv5 with LSTM network. In The International Conference on Intelligent Systems & Networks (pp. 400-408). Singapore: Springer Nature Singapore

[22]Cao, W., Li, L., Gong, S., & Dong, X. (2023, May). Research on human behaviour feature recognition and intelligent early warning methods

in safety supervision scene video based on YOLOv7. Journal of Physics: Conference Series (Vol. 2496, No. 1, p. 012019). IOP Publishing

[23] Human Action Recognition Using CNN-SVM ModelApril 2021Advances in Science and Technology 105:282-290105:282-290DOI:10.4028/www.scientific.net/AST.105.282 Authors:Vijay Anant AthavaleWalchand Institute of Technology Solapur Maharashtra IndiaDeepak KumarState Institute of engineering and technology NilokheriSuresh Chand Gupta