ISSN 2277-2685
IJESR/April-June. 2025/ Vol-15/Issue-2s/515-520
Patnala Pravallika *et. al.*, / International Journal of Engineering & Science Research

Actions detection in video Using Deep Learning

Patnala Pravallika
PG scholar, Department of MCA, DNR College, Bhimavaram, Andhra Pradesh.
K.Suparna
(Assistant Professor), Master of Computer Applications, DNR college, Bhimavaram, Andhra Pradesh.

Abstract: ***The security system gets better with Suspicious Activity Detection using CNNs because this method automatically detects deviations from normal behavior through video analysis. A Convolutional Neural Networks (CNNs) type of deep learning technology enables staffless video surveillance by detecting trespassing and loitering incidents along with aggressive behavior. The security network learns to identify between commonplace and alarming behaviors during its operational period. The system builds its capabilities through video training that combines regular daily activities with security threats. After the training process the model demonstrates the capability to identify security breaches with efficiency. The unique selling point of this solution is its quick processing time combined with high accuracy while operating in real-time. The system functions across diverse environments because the researchers designed it to operate whether light conditions change or cameras are situated differently or background sounds vary. System testing with genuine surveillance video demonstrated its operation success throughout multiple realistic scenarios.***

I. INTRODUCTION

Popular surveillance systems hold a vital position in current-day security protection efforts because safety has risen to unprecedented importance. The majority of security cameras at this time require human intervention for watching extensive video recordings though this method becomes tedious and prone to oversight. Security teams face growing difficulties monitoring vast camera networks and considerable video amounts that require automated smart systems for resolving this task.

The current practice of monitoring surveillance targets through operator visual assessment along with motion tracking and object analysis methods from traditional computer vision remains one of the primary methods. The current monitoring and tracking systems operate too slowly and make mistakes easily whereas their ability to adjust and maintain accuracy remains inadequate. Such systems which operate based on static rules and basic algorithms fail to work efficiently when conditions alter through changes in lighting or camera positioning or unanticipated behavior sequences.

Security protocols become less effective as a result of obtaining incorrect alerts while missing threats. The majority of such systems operate without modern machine learning systems that would enable autonomous learning and adaptation over time. The rigid approach of fixed rules makes these systems ineffective for dealing with unpredictable real-world scenarios thus demonstrating the requirement for automated security improvements.

The artificial intelligence named Convolutional Neural Networks (CNNs) operates exceptionally well for video and image processing tasks. The implementation of CNNs enables automatic surveillance video analysis to identify suspicious

activities which include individuals attempting unauthorized entry into restricted places or remaining in specific areas excessively or displaying confrontational behavior. Real-time alert systems through these systems make security staff work more efficient by providing immediate warnings and enhancing both accuracy and surveillance operations independently of personnel processes.

A CNN-based model learns to identify various suspicious behaviors using real behavioral examples during this project development. During training the system monitors videos containing normal conduct and risky scenarios enabling its ability to identify routine activities from dangerous situations. The combination of automatic learning features with high reliability makes the system capable of effective surveillance for enhancing public safety.

II. LITEARTURE SURVEY

The identification of human actions within video context serves various applications starting from surveillance systems to sports evaluation and interactive human computer systems. Action detection differs from action recognition through its ability to establish both time and space point of occurrence besides identifying recorded movements. A smart model presented here divides human motions into essential poses together with smaller body movements that the researchers call action parts. The system arranges tiny 3D portions (cuboids) surrounding joints into clusters of related movement patterns which it refers to as dynamic-poselets. The poselets enter a skeletal model that learns their behavioral connections throughout space and chronological time for detecting advanced motions in videos. A machine learning technique (structured SVM) within the model trains data to enable quick processing. The method achieved superior performance in processing three popular datasets for action detection which demonstrated its readiness for practical video understanding. [1]

Deep learning has shown great success in tasks like image classification and object detection, but using it effectively for video analysis—such as detecting actions—has been more difficult because of the complexity of video data and the lack of labeled training examples. Most earlier methods detect actions frame by frame and then try to connect those actions across multiple frames, often using separate networks for spatial and motion features. To address these issues, this paper introduces Tube Convolutional Neural Network (T-CNN), an end-to-end deep learning model that detects and recognizes actions directly from videos using 3D convolution. The video is split into smaller clips, action proposals (called “tubes”) are generated for each clip, and then these tubes are linked across clips to detect actions over time and space. Testing on multiple datasets shows that T-CNN outperforms existing methods in accurately identifying and locating actions in both short and long videos. [2]

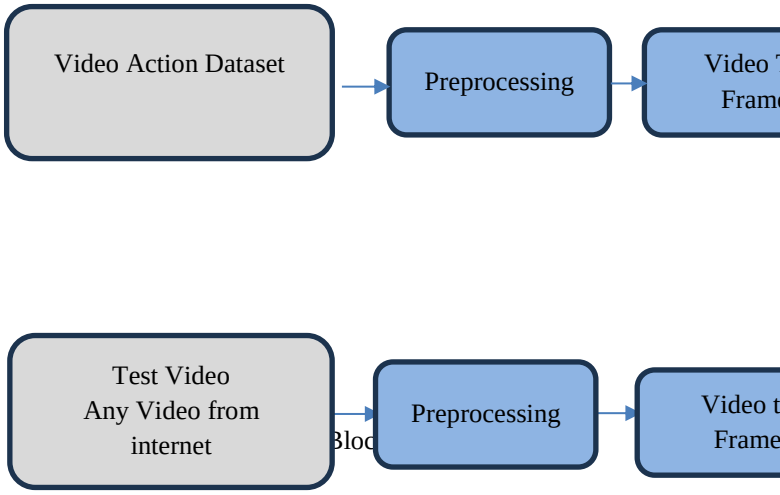
The capability of video-based human action recognition within computer vision applications relies heavily on the chosen features which depict human movements. Many researchers have developed multiple approaches for action detection which use depth data and skeleton information after the Kinect sensor became available. Researchers have not conducted a comprehensive analysis of these methods to assess their feature extraction approaches between human-made techniques and deep learning models along with their reliance on depth images or skeleton data. This paper performs an assessment of ten current Kinect-based action recognition systems which operate under different user and orientation conditions. The analysis demonstrates that most recognition methods

achieve better results with cross-subject deployment and skeleton-based features prove superior to depth features and deep learning techniques deliver optimal outcomes through large training datasets. [3]

Video processing stands as a crucial field in computer vision to construct systems that recognize actions and summarize video content. Logical progress in video action recognition has emerged through deep learning techniques but these approaches commonly examine each of the video frames without addressing numerous repetitive unneeded sequences. The entire frame analysis process leads to waste of time although it results in diminished accuracy levels. A new online system from the authors selects keyframes automatically because these frames provide better action recognition compared to regular frames and improve system speed. The selected keyframes serve as an important tool for both video summarization applications and deep learning algorithms that perform action recognition. The paper presents a new approach which joins keyframes with word vector techniques to enhance system performance. Experimental testing on various widely-used datasets demonstrates this method achieves real-time operation speed while maintaining high accuracy as the first method of its kind. [4]

A novel video recognition architecture named Two-Stream Convolutional Network presents itself as a specific solution for recognizing actions in video content. The model operates two linked streams where spatial information from frames becomes accessible in the spatial stream and temporal motion detection functions in the temporal stream through dense optical flow. The model uses spatial stream appearance analysis together with temporal stream motion recognition to learn complete scene dynamics from video data. The proposed method achieves better results than multiple traditional methods while becoming the new performance standard for video action recognition on UCF-101 and HMDB-51 datasets. The study created fundamental principles which enabled researchers to merge motion and appearance information for deep learning-based video analysis. [5]

The authors in this research document their new deep learning model I3D (Inflated 3D ConvNet) which executes 2D convolutional filters and pooling kernels through inflation processes into 3D for extracting video-based spatiotemporal features. The research proposes Kinetics which serves as a new large-scale human action video dataset that helps advance deeper and more powerful learning of action recognition models despite previous models facing limitations with small datasets. The I3D model reaches best-in-class results across different benchmarks because it proves that 3D convolutional networks trained with enough data produce better outcomes than two-stream approaches. I3D presents an essential advancement in action recognition through deep network architectures trained using abundant datasets. [6]



This research develops a C3D model which uses 3D Convolutional Neural Networks to extract spatiotemporal features from unprocessed video information through temporal convolution extensions from standard 2D methods. Training the C3D network end-to-end through large video datasets leads to effective learning of appearance and motion information between sequential frames. Previous models divided space and time processing functions but C3D merges both operations within one network structure achieving better accuracy with simplified design. Different tasks such as action recognition together with scene

classification and object recognition have been evaluated for strong dataset generalization performance using this model which stands as a powerful video understanding tool. [7]

III. Proposed Method

Proposed method uses python and deep learning for finding actions from the videos.

Proposed method block diagram represent the process of complete proposed system. Proposed system has two main stages

1. Training Stage
2. Testing Stage

Training stage includes below steps,

a) Input Video Action Dataset

In proposed method different types of actions such as shooting , burglary , fighting , boxing , etc are used to train the model.

b) Preprocessing Videos

In video preprocessing , each video is converted to similar size to avoid the loss at training time and to get improved performance.

c) Video to Frame Conversion

Video will be converted to frames and further steps are applied on frames. Video is nothing but sequence of frames.

d) Training the deep learning Model

CNN model is used for training the multiple actions and to apply prediction on new test sample.

e) Fine Tuning and Saving

Fine tuning helps model to understand the data more accuractely.

Testing stage includes below steps,

a) Upload the Test Video

This module will take input from user and feeds to the application. User can select any videos present in dataset or internet videos.

b) Preprocess the video

In video preprocessing , each video is converted to similar size to avoid the loss at training time and to get improved performance.

c) Video to Frame Conversion

Video is nothing but sequence of frames. So each frame from video is used for prediction of action.

d) Prediction Model based on Model Weights

Proposed model after completion of training , weights are getting saved in model folder. The saved model weights are used for further prediction.

e) Predicted Results

Predicted results is an action present in the video , fed by users.

IV. Result Analysis

In proposed system different actions are represented for classification. Convolutional neural network is used for training and testing as below,

Fig. 4. 1 Home page for proposed model

Fig. 4.2Performance of proposed Model

Fig. 4.3 Video Selection

Fig. Upload Video from User to test Anomaly

In above option , user has to change the video to test the action.

Fig. Obtained action is Fighting with probability of matching 0.9956

In above screen in playing video Fighting detected and similarly you can upload

Fig. Video Action Identified as Burglary

In the above video , action identified by proposed CNN model is ‘Burglary’ with probability of 0.9995

In above screen Burglary detected and similarly you can upload and test other videos and while video playing you can press ‘q’ to terminate playing and upload other videos.

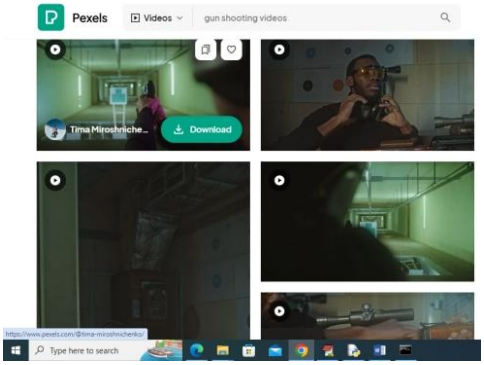


Fig. Collected the video from internet source for testing

Fig. Action detected in Video is shooting

CONCLUSION

The development of VISIONGUARD represents a major advancement for intelligent surveillance by linking AI-based suspicious activity scanning and identity discovery within one consolidated system. The deep learning models along with Convolutional Neural Networks (CNNs) enable real-time identification of abnormal human behavior which then helps security personnel generate prompt responses. The system operates more efficiently with integrated identity detection since it enables proactive threat identification through the recognition of specific individuals. The proposed system decreases the need for human monitoring and eliminates errors while providing faster threat identification during security surveillance of complex crowded settings. The scalable tool VISIONGUARD operates with high efficiency to support surveillance operations across public safety measures and smart cities along with transportation hubs as well as secure facilities. The system will be improved through additions like multi-camera coordination features and front-facing recognition technology as well as deployment mechanisms for distributed edge device processing capabilities.

REFERENCES

1. Wang, Limin, Yu Qiao, and Xiaoou Tang. "Video action detection with relational dynamic-poselets." In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 565-580. Springer International Publishing, 2014.
2. Hou, Rui, Chen Chen, and Mubarak Shah. "Tube convolutional neural network (t-cnn) for action detection in videos." In *Proceedings of the IEEE international conference on computer vision*, pp. 5822-5831. 2017.
3. Wang, Lei, Du Q. Huynh, and Piotr Koniusz. "A comparative review of recent kinect-based action recognition algorithms." *IEEE Transactions on Image Processing* 29 (2019): 15-28.
4. Elahi, GM Mashrur E., and Yee-Hong Yang. "Online learnable keyframe extraction in videos and its application with semantic word vector in action recognition." *Pattern Recognition* 122 (2022): 108273.
5. Simonyan, K., & Zisserman, A. (2014). "Two-stream convolutional networks for action recognition in videos." Advances in Neural Information Processing Systems (NeurIPS).
6. Carreira, J., & Zisserman, A. (2017). "Quo vadis, action recognition? A new model and the kinetics dataset." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
7. Tran, D. et al. (2015). "Learning spatiotemporal features with 3D convolutional networks." International Conference on Computer Vision (ICCV).
8. Girdhar, R. et al. (2018). "Detect-and-track: Efficient pose-based action recognition in videos." IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
9. Wang, L. et al. (2016). "Temporal segment networks: Towards good practices for deep action recognition." European Conference on Computer Vision (ECCV).
10. Sun, C. et al. (2019). "VideoBERT: A joint model for video and language representation learning." IEEE International Conference on Computer Vision (ICCV).
11. Gavrilyuk, K. et al. (2020). "Actor-transformer: A transformer architecture for action localization." British Machine Vision Conference (BMVC).
12. Feichtenhofer, C. et al. (2019). "SlowFast networks for video recognition." IEEE International Conference on Computer Vision (ICCV).
13. Arnab, A. et al. (2021). "ViViT: A Video Vision Transformer." International Conference on Computer Vision (ICCV).
14. Bertasius, G. et al. (2021). "Space-time attention for efficient video understanding." International Conference on Machine Learning (ICML).