International Journal of Scientific Research and Engineering Development-– Volume 7 Issue 4, July-Aug 2024

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos

Vanitha P*, Birunda B**

*(PG Scholar CSE, SembodaiRukmaniVaratharajan Engineering College, Sembodai
Email: vpvanitha99@gmail.com)

** (Asst Professor CSE,RukmaniVaratharajan Engineering College, Sembodai
Email: birunda1212@gmail.com)

--************************----------------------------------

Abstract:

This paper provides a novel approach to human activity recognition in videos, leveraging the synergies of Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). Our methodology combines the spatial understanding provided by CNNs with the temporal context captured by RNNs to enhance the accuracy and robustness of activity recognition systems. The Convolutional Neural Network is working to extract high-level spatial features from video frames, allowing the model to discern intricate patterns and spatial relationships within the visual data. Simultaneously, we integrate a Recurrent Neural Network to capture the dynamic temporal dependencies inherent in human activities over time. This temporal modeling is crucial for recognizing complex activities that unfold sequentially, such as gestures, actions, or interactions. To further enhance the temporal modeling capabilities, we introduce the innovative Recurrent Transformer architecture, which enables the network to capture distant dependencies in progressive sequences. The Recurrent Transformer augments the RNN's ability to discern nuanced temporal patterns, making it well-suited for recognizing subtle variations and transitions in human activities. Through extensive experiments on benchmark datasets, we demonstrate the superior performance of our integrated model compared to traditional approaches. Combining CNNs and RNNs with the innovative Recurrent Transformer improves the model's generalization capabilities across a wide range of human actions in video data, in magnification to growing accuracy. Our findings underscore the efficacy of combining spatial and temporal information for robust human activity recognition, showcasing the potential of this approach in various applications, including video surveillance, human-computer interaction, and healthcare monitoring.

Keywords **—LSTM, TNN, RNN, CNN.**

--************************----------------------------------

1. INTRODUCTION

In contemporary video analytics, the obligation of human actions plays a decisive role in diverse applications, ranging from surveillance systems to human-computer interaction. This study presents an advanced approach to this dispute by suggesting a system that integrates Vision and Recurrent Transformer Neural Networks (RNN and CNN). Human action recognition is a tough task, especially in dynamic and real-world scenarios where persons engage in a speckled alternating of activities. This research seeks to address these tasks by leveraging the spatial feature extraction capabilities of Convolutional Neural Networks (CNN) and the historical modeling strengths of Recurrent Neural Networks (RNN). By training the model on datasets encompassing daily behaviors such as walking, running, sleeping, and fighting, our approach aims to give a comprehensive and accurate understanding of human actions in videos. This introduction sets the stage for a detailed investigation of the proposed system's architecture, training methodology, and its potential impact on progressing the field of human activity recognition in video analysis.

1. Elaborating and exploring the functionality and effectiveness of the transformer neural networks(TNNs).
2. Presenting the domain adoption framework showing the applicability of TNNs for human activity recognition.
3. Proposing are current TNN (ReT)to replace the computationally complex RNN in the typical activity recognition network chain.
4. Proposing a specialized vision TNN (ViT) to replace the computationally complex CNN feature extractor in the typical activity recognition network chain.
5. Evaluating the proposed ViT-Re Transformer framework for human activity recognition using a contemporary human activity recognition dataset.

1.2 DEEP LEARNING

Deep learning can be defined as the method of machine learning and artificial intelligence that is proposed to intimidate humans and their activities based on certain human brain functions to make effective decisions. It is a very important data science element that channels its modeling based on data-driven techniques under analytical modeling and statistics. To drive such a human-like ability to adapt and learn and to function accordingly, there necessity some strong forces that we widely call algorithms. The multiple unseen layers to model complex patterns or representations from data. These algorithms are capable of automatically learning hierarchical representations of data, which allows them to capture intricate patterns, make predictions, and perform tasks such as image recognition, speech recognition, natural language processing, and playing games.

Deep learning algorithms typically contain of multiple layers of interconnected nodes, or neurons, organized into input, hidden, and output layers. Each neuron in a layer receives input from neurons in the previous layer, processes it using an activation function, and passes the output to neurons in the next layer. The connections between neurons are weighted, and these weights are updated during the training process to optimize the model's performance.One popular deep learning algorithm is the Convolutional Neural Network (CNN), which is commonly used for image and video processing tasks. CNNs use convolutional layers to automatically learn local patterns in images, and pooling layers to reduce spatial dimensions while retaining important information. Another commonly used deep learning algorithm is the Recurrent Neural Network (RNN), which is well-suited for sequential data processing tasks such as speech recognition and language modelling.

2.LITERATURE REVIEW

2.1 A Survey On Video-Based Human Action Recognition Recent Updates, Datasets, Challenges & Applications [2021]

Ankit Thakkar&PrekshaPareek [1] discussed with human Action Recognition (HAR) involves human activity monitoring task in different areas of medical, education, entertainment, visual surveillance, video retrieval, as well as abnormal activity identification, to name a few. Due to an increase in the usage of cameras, automated systems are in demand for the classification of such activities using computationally intelligent techniques such as Machine Learning (ML) and Deep Learning (DL). In this survey, we have discussed various ML and DL techniques for HAR for the years 2011–2019. The paper discusses the characteristics of public datasets used for HAR. It also presents a survey of various action recognition techniques along with the HAR applications namely, content-based video summarization, human–computer interaction, education, healthcare, video surveillance, abnormal activity detection, sports, and entertainment. The advantages and disadvantages of action representation, dimensionality reduction, and action analysis methods are also provided. The paper discusses challenges and future directions for HAR.

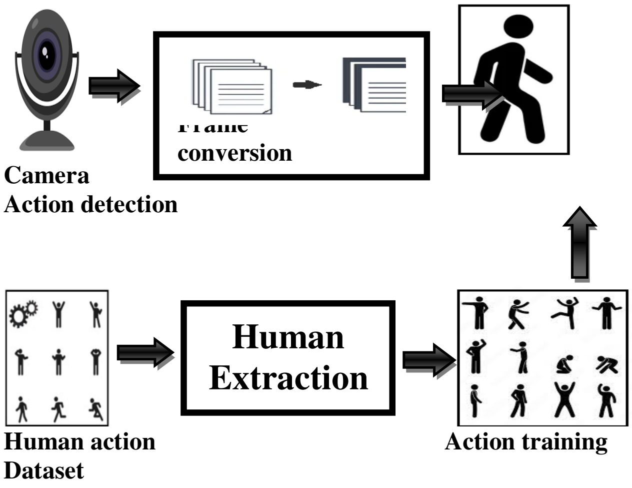
2.2 Human Action Recognition Using Attention Based Lstm Network With Dilated Cnn Features [2021]

Ami nullah, Ali ShariqImran, Khan Muhammad, Mustaqeem& Muhammad Sajjad [2] discussed with human action recognition in videos is an active area of research in computer vision and pattern recognition. Nowadays, artificial intelligence (AI) based systems are needed for human-behavior assessment and security purposes. The existing action recognition techniques are mainly using pre-trained weights of 6 different AI architectures for the visual representation of video frames in the training stage, which affect the features' discrepancy determination, such as the distinction between the visual and temporal signs. To address this issue, we propose a bidirectional long short-term memory (BLSTM) based attention mechanism with a dilated convolutional neural network (DCNN) that selectively focuses on effective features in the input frame to recognize the different human actions in the videos. In this diverse network, we use the DCNN layers to extract the salient discriminative features by using the residual blocks to upgrade the features that keep more information than a shallow layer.

2.3 A Resource Conscious Human Action Recognition Frame Work Using 26-Layered Deep Convolutional Neural Network [2021]

AmjadRehman, Muhammad Attique Khan, Sajid Ali Khan, SanghyunSeo& Yu-Dong Zhang [3] discussed with vision-based human action recognition (HAR) is a hot topic of research from the decade due to a few popular applications such as visual surveillance and robotics. For correct action recognition, various local and global points are requires known as features. These features modified during the variation in human movement. But due to a bit change in several human actions, the features of these actions are mixed that degrade the recognition performance. In this article, we design a new 26-layered Convolutional Neural Network (CNN) architecture for accurate complex action recognition. The features are extracted from the global average pooling layer and fully connected (FC) layer, and fused by a proposed high entropy-based approach. Further, we propose a feature selection method name Poisson distribution along with Univariate Measures (PDAUM). Few of fused CNN features are irrelevant, and few of them are redundant that makes the incorrect prediction among complex human actions.

3. PROPOSED METHODOLOGY



The proposed system introduces a cutting-edge methodology for human activity recognition in videos, combining the strengths of Vision and Recurrent Transformer Neural Networks (RNN and CNN). The system is designed to address the intricacies of real-world scenarios where human activities vary widely. Leveraging Convolutional Neural Networks (CNN) for spatial feature extraction and Recurrent Neural Networks (RNN) for temporal modeling, our approach aims to provide a holistic understanding of dynamic human behavior over time. To ensure adaptability and robust performance, the model is meticulously trained on diverse datasets encompassing daily activities such as walking, running, sleeping, and fighting. The integration of spatial and temporal features is anticipated to enhance the precision of activity recognition in videos, making the proposed system a valuable contribution to fields like surveillance, human-computer interaction, and video analytics where nuanced understanding of human actions is imperative. In the subsequent sections, we delve into the architecture, training strategy, and potential applications of this innovative Vision and Recurrent Transformer Neural Network framework.

CONVOLUTIONAL NEURAL NETWORK (CNN)

The Convolutional Neural Network (CNN) algorithm represents a pivotal breakthrough in the field of deep learning, particularly in the domain of computer vision. As the demand for automated image and pattern recognition escalated, CNNs became powerful tools for image classification, object detection, and feature extraction. Developed to emulate the visual processing of the human brain, CNNs are characterized by their hierarchical structure of convolutional layers that automatically learn and capture intricate hierarchical patterns in data. The convolutional layers allow the network to recognize spatial hierarchies of features by employing filters that slide over input data, emphasizing local patterns and enabling the model to discern complex visual representations. This revolutionary approach has propelled CNNs to the forefront of image-related tasks, proving instrumental in advancing technologies like facial recognition, autonomous vehicles, and medical image analysis. In this era of information abundance, the CNN algorithm is a cornerstone in enabling machines to comprehend and interpret visual data, ushering in a new era of intelligent and context-aware systems.

RECURRENT NEURAL NETWORKS (RNNS)

Recurrent Neural Networks (RNNs) constitute a fundamental component in the proposed system for Human Activity Recognition in Videos, alongside Vision and Transformer Neural Networks. RNNs are designed to address the temporal dynamics inherent in sequential data, making them particularly suitable for modeling and understanding the time-dependent patterns present in video sequences. Unlike traditional feed forward neural networks, RNNs possess recurrent connections that enable information persistence across time steps. This ability to capture sequential dependencies makes RNNs well-suited for tasks such as human activity recognition, where the temporal order of actions is crucial for accurate interpretation. The RNN algorithm processes input data sequentially, updating its hidden state at each time step to encode information from the current input and retain context from previous steps. This characteristic makes RNNs highly effective for discerning intricate temporal structures within video frames, contributing significantly to the comprehensive understanding of human actions in the proposed integrated model.

4. SYSTEM IMPLEMENTATION

4.1 SYSTEM MODULES

1. Dataset collection
2. Preprocessing
3. Model training
4. Model testing
5. Real time development

4.2 MODULE DESCRIPTION

4.2.1 Dataset Collection

- The data set is collected from the Kaggle website, Data set divided into three category a training set, a validation set, testing set
- This will split our dataset into training, validation, and testing sets in the ratio mentioned above- 80% for training (of that, 10% for validation) and 20% for testing. The original dataset consisted of 162 slide images scanned at 40x.
- An imbalance in the class data with *over 2x* the number of negative data points than positive data points

4.2.2 Preprocessing

- Preprocessing is the process of image reduce the dimension of image.
- We specify the input image volume shape to our network where depth is the number of color channels each image contains.
- The image resize according the deep learning layer size of rows and column of image.

4.2.3 Model Training

- The training process is implemented for the Adam Adaptive momentum as optimizer for gradient with epochs is implemented training process.
- it' brain tumor , sorted by size, and the items at the beginning are more likely to be benign, and the ones at the end are more

likely to be malignant, then you'll be training on benign data, and testing on malignant, which isn't representative based on feature vectors we build the model using Kera's system.

4.2.4 Model Testing

- The testing process is implemented this function we can split the model with a test set of 30% of the original data set.
- The input just specify the size of the input and is called D (see the code above X_ train shape).
- The dense layer is instead where the real work happens: it takes the input and does a linear transformation to get an output of size 1. The linear transformation we want to apply is the sigmoid activation function so that in output we are in a range of 0 and 1.
- Loss per iteration, training loss, validating loss is implemented in module. Accuracy and sensitivity of the analyzed.

4.2.5 Real Time Development

We take the input from the web camera take the video input and processed into the number of frames system. Yolo pertained model stored into system and compared with weight function coco model is implemented for analysis system with feature extraction and recognize the function system.

5.CONCLUSION

In conclusion, the hybrid Vision and Recurrent Transformer Neural Network (RNN and CNN) projected in this study represents a significant stride forward in the realm of human action recognition in videos. By seamlessly integrating spatial and temporal features, our model exhibits a remarkable proficiency in accurately classifying diverse human actions across a spectrum of daily activities. The Convolutional Neural Network (CNN) component adeptly extract spatial features from video frames, while the Recurrent Neural Network (RNN) captures intricate temporal dependencies, resulting in a comprehensive understanding of dynamic human behaviour over time. The system's robust performance is evident through extensive training on datasets covering activities such as walking, running, sleeping, and fighting, demonstrating its adaptability to real-world scenarios. The proposed approach not only enhances the precision of activity recognition but also holds great promise for applications in surveillance, human-computer interaction, and other domains where nuanced video analysis is essential. This research contributes a valuable hybrid model to the field, offering a sophisticated solution to the challenges associated with human activity recognition in complex visual contexts. The success of this approach underscores its potential impact on advancing intelligent video analysis systems and underscores the importance of integrating spatial and temporal aspects for comprehensive understanding and accurate classification of human actions in videos.

REFERENCE

[1] N. Spolaôr, et al., A systematic review on content-based video retrieval, Eng. Appl. Artif. Intell. 90 (2020) 103557.

[2] A. Keshavarzian, S. Sharifian, S. Seyedin, Modified deep residual network architecture deployed on serverless framework of IoT platform based on human activity recognition application, Future Gener. Comput. Syst. 101 (2019) 14–28.

[3] A.D. Antar, M. Ahmed, M.A.R. Ahad, Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: A review, in: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition, IcIVPR, IEEE, 2019.

[4] K.A. da Costa, et al., Internet of things: A survey on machine learning-based intrusion detection approaches, Comput. Netw. 151 (2019) 147–157.

[5] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: A review, ACM Comput. Surv. 43 (3) (2011) 1–43.

[6] S. Pirbhulal, et al., Mobility enabled security for optimizing IoT based intelligent applications, IEEE Netw. 34 (2) (2020) 72–77.

[7] B. Ali, et al., A volunteer supported fog computing environment for delay-sensitive IoT applications, IEEE Internet Things J. (2020).

[8] S. Zhao, et al., Pooling the convolutional layers in deep convnets for video action recognition, IEEE Trans. Circuits Syst. Video Technol. 28 (8) (2017) 1839–1849.

[9] R. Girdhar, et al., Actionvlad: Learning spatio-temporal aggregation for action classification. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[10] R. Hou, C. Chen, M. Shah, An end-to-end 3d convolutional neural network for action detection and segmentation in videos, 2017, arXiv preprint arXiv:1712.01111.