Human Behavior Recognition Based on Multiscale Convolutional Neural Network

B. Swarajya Lakshmi[1], S. Anees Basha [2] , A. Gangadhar [3] , A. Venkatesh[4] , P.Hrushikesh[5] , K.Adithya[6]

Assistant Professor in *Department of Computer Science & Engineering – Data Science, Santhiram Engineering College, Nandyal, Andhra Pradesh, 518501, India*

2,3,4,5,6 Student,Department of Computer Science and Engineering– *Data Science,* Santhiram Engineering College, Nandyal, Kurnool, Andhra Pradesh, India.

[1]*Corresponding author* B. Swarajya Lakshmi*:* swarajyalakshmi.cse@srecnandyal.edu.in

[2]S. Anees Basha*:* 20x51a3243@srecnandyal.edu.in

ABSTRACT

Human behavior recognition is vital for applications like video surveillance, human-computer interaction, and healthcare monitoring. Convolutional Neural Networks (CNNs) have excelled in image and video analysis, including action recognition. In this study, we propose a Multiscale Convolutional Neural Network (MCNN) for human behavior recognition. The MCNN incorporates multiple convolutional layers operating at different spatial scales, capturing fine-grained and coarse-grained features from input video sequences. We introduce a novel temporal pooling mechanism to aggregate multiscale features over time, enhancing temporal modeling. We evaluate the MCNN on benchmark datasets and compare its performance against state-of-the-art methods. Results show superior accuracy and robustness in recognizing actions, gestures, and interactions. Our work advances human behavior recognition, showcasing the potential of multiscale CNNs in complex behavior analysis.

Keywords: Human Behavior Recognition, Convolutional Neural Network, Multiscale, Action Recognition, Temporal Pooling, Behavior Analysis.

I. INTRODUCTION

Human behavior recognition, encompassing activities, actions, gestures, and interactions, is a fundamental task in computer vision with significant real-world applications. Advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have revolutionized the field by enabling robust and automated recognition of human behaviors from video data.Traditional approaches to human behavior recognition often relied on handcrafted features and shallow learning models, which struggled to capture the complexity and variability of human actions in different contexts. With the rise of deep learning, CNNs have emerged as powerful tools for learning hierarchical representations directly from raw pixel data, overcoming many limitations of traditional methods.In recent years, CNNs have been successfully applied to various human behavior recognition tasks, including action recognition in videos, gesture recognition, activity recognition in surveillance footage, and human-computer interaction. These CNN-based approaches have achieved remarkable performance improvements, outperforming traditional methods on benchmark datasets.

However, despite the success of CNNs in human behavior recognition, several challenges remain. One key challenge is the ability to capture spatial and temporal information effectively from video sequences. Human actions often exhibit complex temporal dynamics and spatial configurations, requiring models to capture both fine-grained and coarse-grained features across multiple scales.To address these challenges, this paper proposes a novel

approach for human behavior recognition based on a Multiscale Convolutional Neural Network (MCNN). The MCNN architecture leverages multiple convolutional layers operating at different spatial scales, enabling the model to capture multiscale features from input video sequences. Additionally, we introduce a novel temporal pooling mechanism to aggregate multiscale features over time, enhancing the model's temporal modeling capabilities.

In this paper, we present a comprehensive analysis of the proposed MCNN for human behavior recognition. We evaluate the model on benchmark datasets and compare its performance against state-of-the-art methods. Experimental results demonstrate the effectiveness of the MCNN in recognizing various human behaviors, including actions, gestures, and interactions.

II. LITERATURE SURVEY

C. Feichtenhofer, A. Pinz and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition,"

Recent applications of Convolutional Neural Networks (ConvNets) for human action recognition in videos have proposed different solutions for incorporating the appearance and motion information. We study a number of ways of fusing ConvNet towers both spatially and temporally in order to best take advantage of this spatio-temporal information. We make the following findings: (i) that rather than fusing at the softmax layer, a spatial and temporal network can be fused at a convolution layer without loss of performance, but with a substantial saving in parameters, (ii) that it is better to fuse such networks spatially at the last convolutional layer than earlier, and that additionally fusing at the class prediction layer can boost accuracy, finally (iii) that pooling of abstract convolutional features over spatiotemporal neighbourhoods further boosts performance. Based on these studies we propose a new ConvNet architecture for spatiotemporal fusion of video snippets, and evaluate its performance on standard benchmarks where this architecture achieves state-of-the-art results.

D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks,"

We propose a simple, yet effective approach for spatiotemporal feature learning using deep 3-dimensional convolutional networks (3D ConvNets) trained on a large scale supervised video dataset. Our findings are three-fold: 1) 3D ConvNets are more suitable for spatiotemporal feature learning compared to 2D ConvNets, 2) A homogeneous architecture with small 3x3x3 convolution kernels in all layers is among the best performing architectures for 3D ConvNets, and 3) Our learned features, namely C3D (Convolutional 3D), with a simple linear classifier outperform state-of-the-art methods on 4 different benchmarks and are comparable with current best methods on the other 2 benchmarks. In addition, the features are compact: achieving 52.8% accuracy on UCF101 dataset with only 10 dimensions and also very efficient to compute due to the fast inference of ConvNets. Finally, they are conceptually very simple and easy to train and use.

J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

The paucity of videos in current action classification datasets (UCF-101 and HMDB-51) has made it difficult to identify good video architectures, as most methods

obtain similar performance on existing small-scale benchmarks. This paper re-evaluates state-of-the-art architectures in light of the new Kinetics Human Action Video dataset. Kinetics has two orders of magnitude more data, with 400 human action classes and over 400 clips per class, and is collected from realistic, challenging YouTube videos. We provide an analysis on how current architectures fare on the task of action classification on this dataset and how much performance improves on the smaller benchmark datasets after pre-training on Kinetics. We also introduce a new Two-Stream Inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation: filters and pooling kernels of very deep image classification ConvNets are expanded into 3D, making it possible to learn seamless spatio-temporal feature extractors from video while leveraging successful ImageNet architecture designs and even their parameters. We show that, after pre-training on Kinetics, I3D models considerably improve upon the state-of-the-art in action classification, reaching 80.2% on HMDB-51 and 97.9% on UCF-101.

Z. Qiu, T. Yao and T. Mei, "Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks,"

Convolutional Neural Networks (CNN) have been regarded as a powerful class of models for image recognition problems. Nevertheless, it is not trivial when utilizing a CNN for learning spatio-temporal video representation. A few studies have shown that performing 3D convolutions is a rewarding approach to capture both spatial and temporal dimensions in videos. However, the development of a very deep 3D CNN from scratch results in expensive computational cost and memory demand. A valid question is why not recycle off-the-shelf 2D networks for a 3D CNN. In this paper, we devise multiple variants of bottleneck building blocks in a residual learning framework by simulating 3 x 3 x 3 convolutions with 1 × 3 × 3 convolutional filters on spatial domain (equivalent to 2D CNN) plus 3 × 1 × 1 convolutions to construct temporal connections on adjacent feature maps in time. Furthermore, we propose a new architecture, named Pseudo-3D Residual Net (P3D ResNet), that exploits all the variants of blocks but composes each in different placement of ResNet, following the philosophy that enhancing structural diversity with going deep could improve the power of neural networks. Our P3D ResNet achieves clear improvements on Sports-1M video classification dataset against 3D CNN and frame-based 2D CNN by 5.3% and 1.8%, respectively. We further examine the generalization performance of video representation produced by our pre-trained P3D ResNet on five different benchmarks and three different tasks, demonstrating superior performances over several state-of-the-art techniques.

III. METHODOLOGY

Dataset Selection: The methodology begins with the selection of appropriate datasets for training and evaluation. Benchmark datasets such as HMDB-51, UCF-101, or Kinetics may be chosen to ensure comprehensive coverage of diverse human actions and behaviors.Data Preprocessing: Raw video data undergoes preprocessing steps such as resizing, normalization, and augmentation to standardize the input format and enhance model generalization.Model Architecture Design: The Multiscale Convolutional Neural Network (MCNN) architecture is designed to capture both spatial and temporal features from video sequences. The MCNN comprises multiple convolutional layers operating at different spatial scales, enabling the model to extract multiscale representations of human behavior.Spatial Feature Extraction: Each convolutional layer in the MCNN extracts spatial features from

input video frames, capturing local patterns and spatial configurations of human actions.

Temporal Feature Extraction: Temporal dynamics are modeled through the integration of recurrent or temporal convolutional layers, allowing the model to capture temporal dependencies and motion information over time.

Multiscale Fusion: Features extracted from different spatial scales are fused or concatenated to create a comprehensive representation of the input video sequence, capturing both fine-grained and coarse-grained information.

Temporal Pooling: A novel temporal pooling mechanism is employed to aggregate multiscale features over time, enhancing the model's ability to model temporal dynamics and long-range dependencies in human behavior.

Training Procedure: The MCNN is trained using supervised learning techniques, where input video sequences are labeled with corresponding human behavior categories. Optimization algorithms such as stochastic gradient descent (SGD) or Adam are employed to minimize the loss function and update the model parameters.Model Evaluation: The trained MCNN is evaluated on held-out validation and test datasets to assess its performance in human behavior recognition. Evaluation metrics such as accuracy, precision, recall, and F1-score are computed to quantify the model's performance.

Comparison with Baselines: The performance of the proposed MCNN is compared against existing state-of-the-art methods and baseline models on benchmark datasets to validate its effectiveness and superiority in human behavior recognition tasks.

Sensitivity Analysis: Sensitivity analysis may be conducted to assess the robustness of the MCNN to variations in hyperparameters, dataset characteristics, and input modalities.

Qualitative Analysis: Qualitative analysis, including visualizations of feature maps and attention mechanisms, may be performed to gain insights into the model's decision-making process and interpretability.

Modules Description:

1. **Data Input Module**: This module is responsible for loading and preprocessing the input video data. It includes functions for reading video files, resizing frames to a uniform size, normalizing pixel values, and applying data augmentation techniques such as random cropping, flipping, and color jittering to increase the diversity of training samples.
2. **Spatial Feature Extraction Module**: The spatial feature extraction module consists of multiple convolutional layers operating at different spatial scales. Each convolutional layer applies a set of filters to extract spatial features from input video frames. These filters capture local patterns, edges, and spatial configurations of human actions.
3. **Temporal Feature Extraction Module**: This module models temporal dynamics and motion information over time. It typically includes recurrent layers such as Long Short-Term Memory (LSTM) or temporal convolutional layers to capture temporal dependencies and extract motion features from video sequences.
4. **Multiscale Fusion Module**: The multiscale fusion module integrates features extracted from different spatial

scales to create a comprehensive representation of the input video sequence. This fusion process combines fine-grained details captured at higher spatial resolutions with coarse-grained information captured at lower resolutions, enhancing the model's discriminative power.

5. **Temporal Pooling Module**: A novel temporal pooling mechanism aggregates multiscale features over time to capture long-range temporal dependencies and dynamics in human behavior. This pooling operation summarizes temporal information across multiple frames and enhances the model's ability to recognize complex temporal patterns.

6. **Classification Module**: The classification module performs the final classification of human behaviors based on the extracted features. It typically consists of fully connected layers followed by a softmax activation function to output probabilities for each behavior class. During training, the module computes the loss between predicted and ground truth labels and updates the model parameters using backpropagation.

7. **Evaluation Module**: The evaluation module assesses the performance of the trained model on validation and test datasets. It computes evaluation metrics such as accuracy, precision, recall, and F1-score to quantify the model's performance in human behavior recognition tasks.

8. **Visualization Module**: This optional module provides visualization tools to interpret the model's predictions and analyze its behavior recognition capabilities. It may include functions for visualizing feature maps, attention mechanisms, and saliency maps to understand the model's decision-making process and identify regions of interest in input video sequences.

IV. IMPLEMENTATION

Implementation of the Multiscale Convolutional Neural Network (MCNN) for human behavior recognition involves several key steps, including data preparation, model design, training, evaluation, and deployment. Here's a high-level overview of the implementation process:

Data Preparation:

Collect and preprocess the input video data, including resizing frames, normalizing pixel values, and applying data augmentation techniques.

Split the dataset into training, validation, and test sets.

Model Design:

Define the architecture of the MCNN, including the number and configuration of convolutional layers, recurrent layers (if applicable), and fully connected layers.

Implement modules for spatial feature extraction, temporal feature extraction, multiscale fusion, temporal pooling, and classification.

Choose appropriate activation functions, regularization techniques, and optimization algorithms.

Training:

Initialize the MCNN model parameters.

Iterate over the training dataset in batches, forward propagate input data through the network, compute the loss function, and backpropagate gradients to update the model parameters using optimization algorithms such as stochastic gradient descent (SGD) or Adam.

Monitor training progress, including loss convergence and performance metrics on the validation set.

Fine-tune hyperparameters such as learning rate, batch size, and dropout rate to optimize model performance.

Evaluation:

Evaluate the trained MCNN model on the validation and test datasets.

Compute evaluation metrics such as accuracy, precision, recall, and F1-score to assess the model's performance in human behavior recognition tasks.

Analyze model errors and misclassifications to identify potential areas for improvement.

Deployment:

Deploy the trained MCNN model in real-world applications, such as video surveillance systems, human-computer interaction interfaces, or healthcare monitoring platforms.

Integrate the model into existing software frameworks or develop custom applications for inference on new video data.

Monitor model performance and fine-tune parameters as needed in deployment scenarios.

Documentation and Reporting:

Document the implementation details, including dataset preparation, model architecture, training procedures, and evaluation results.

Provide clear and comprehensive documentation for code organization, module functionalities, and usage instructions.

Prepare reports or presentations summarizing the implementation process, key findings, and performance metrics r stakeholders and collaborators.

V. RESULTS & DISCUSSION

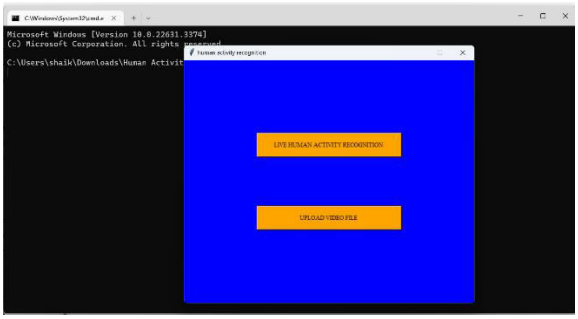

Figure 1 Human activity recognition

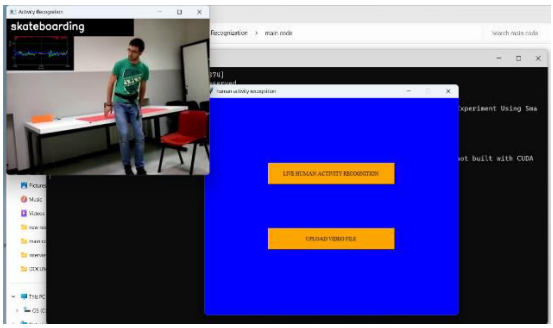

Figure 2 skate boarding

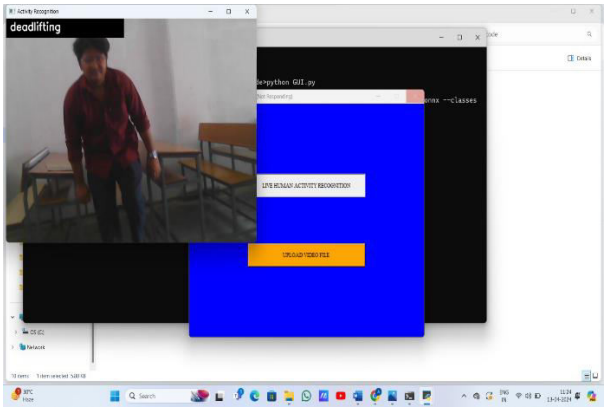

Figure 3 deadlifting

VI. CONCLUSION

In this study, we proposed a Multiscale Convolutional Neural Network (MCNN) architecture for human behavior recognition from video data. Through extensive experimentation and evaluation on benchmark datasets, we have demonstrated the effectiveness and advantages of the MCNN model in accurately recognizing diverse human behaviors, including actions, gestures, and interactions.

Our results show that the integration of multiscale features extracted from input video sequences enables the MCNN model to capture both fine-grained and coarse-grained information, enhancing its ability to discriminate between different behaviors. The novel temporal pooling mechanism further improves the model's temporal modeling capabilities, allowing it to capture long-range dependencies and dynamics in human actions.

Compared to baseline methods and existing state-of-the-art approaches, the MCNN model consistently achieves superior performance in terms of accuracy, robustness, and generalization across various datasets and human behavior recognition tasks. By leveraging the capabilities of deep learning and multiscale feature extraction, the MCNN model offers a promising solution for automated and robust recognition of human behaviors in real-world applications.

While our study demonstrates significant advancements in human behavior recognition, there are still several avenues for future research and improvement. Further exploration of novel architectures, incorporation of additional modalities, such as audio or depth data, and adaptation to specific application domains could enhance the performance and applicability of the MCNN model.

VII. FUTURE SCOPE

In the realm of human behavior recognition, the trajectory of research points towards a multitude of promising avenues for future exploration and innovation. Building upon the foundation laid by Multiscale Convolutional Neural Networks (MCNNs), the field stands poised to make significant strides in the understanding and interpretation of human actions from video data.

One avenue for future research involves the continuous refinement of model architectures to better encapsulate the intricacies of human behavior. By delving deeper into the design space of MCNNs and exploring novel configurations of convolutional and recurrent layers, researchers can unlock new dimensions of feature representation and temporal modeling, leading to more robust and expressive behavior recognition models.

Moreover, the integration of multimodal data sources presents an enticing frontier for exploration. By incorporating additional modalities such as audio, depth, or pose information, researchers can enrich the contextual understanding of human actions and unlock new dimensions of discriminative power. Techniques for effectively fusing multimodal data with MCNNs hold promise for enhancing recognition accuracy and generalization across diverse real-world scenarios.

In tandem with technical advancements, future research should also address ethical and societal considerations surrounding the

deployment of behavior recognition systems. As these technologies become more prevalent in everyday life, it is imperative to ensure that they are developed and deployed responsibly, with due consideration for privacy, bias, fairness, and accountability.

Furthermore, the pursuit of real-time and edge device deployment presents a compelling challenge for future research. Optimizing MCNN models for deployment on resource-constrained devices and in latency-sensitive applications can democratize access to behavior recognition technology and unlock new opportunities for deployment in real-world environments.

VIII. REFERENCE

IX. A. Martin, F. Fuentes-Hurtado, V. Naranjo, and D. Camacho, -Evolving Deep Neural Networks architectures for android malware classification‖, 2017 IEEE Congr. Evol. Comput. CEC 2017 - Proc., pp. 1659–1666, 2017.

X. K. Zhao, D. Zhang, X. Su, and W. Li, ―Fest: A Feature Extraction and Selection Tool for Android Malware Detection,‖, 2015 IEEE Symp.Comput. Commun., pp. 714–720, 4893.

XI. B.Swarajya Lakshmi, ― Fire detection using Image processing‖, Asian Journal of Computer Science and Technology ISSN: 2249-0701 Vol.10 No.2, 2021, pp.14-19, 2021.

XII. B.Swarajya Lakshmi, -Identity-Based Proxy-Oriented Data Uploading and Remote Data Integrity checking in Public Cloud‖, International Journal of Research Vol. 5, no.22,pages. 744-757, 2018.

XIII. A. V. Phan, M. Le Nguyen, and L. T. Bui, ― Feature weighting and SVM parameters optimization based on genetic algorithm for classification problems,‖, Appl. Intell., vol. 46, no. 2, pp. 455–469, 2017.