Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 7
July 2023

Human Action Recognition using an Ensemble Deep Learning Model for Video Datasets

Ramanpreet Kaur[1], Dr. Dharam Veer Sharma[2]
[1,2]Department of Computer Science, Punjabi University, Patiala, India
ramanpreet.star@gmail.com, dveer72@hotmail.com

Abstract: Human Action Recognition is used to analyse the videos to identify the actions performed by humans. In recent years, it has gained much popularity due to its large domain of applications presented in various fields. Several research contributions are available in this area but still the requirement is to achieve good results for various challenging datasets and limited hardware resources. In order to overcome these issues, an ensemble deep learning model is proposed in this paper based on custom convolutional neural network (CNN). ResNet50 is merged with a handcrafted CNN, to identify human actions in challenging video datasets. This model is trained and tested on UCF-101 and HMDB-51 datasets and gained very good results. The experimental results showcased that the proposed model outperforms some recent works in this domain.
Keywords: HAR, CNN, ResNet50, Deep Learning

1. Introduction

Human Action or Activity Recognition involves identifying the goals or objectives of one or more actions by analysing a series of observed actions and their environmental conditions. Achieving accurate activity recognition poses various challenges due to the complexity and diversity of human activities. The terms "action" and "activity" are often used interchangeably. An action refers to a fundamental human motion, such as jumping or walking. On the other hand, an activity is a complex human motion that comprises multiple basic actions. For example, drinking water can be defined as a sequence of actions: opening a bottle, pouring water into a glass, and then drinking it [1]. The recognition of human actions from videos and images has garnered significant attention over the past few decades. This field has numerous practical applications in daily life due to the widespread use of cameras. Surveillance cameras aid in fraud detection, human-computer interaction enables gaming experiences, and activity detection assists in healthcare and elder care. Additionally, activity recognition plays a vital role in content-based video retrieval, as most web search engines heavily rely on textual data. Over the years, researchers have proposed numerous techniques and approaches for human action and activity recognition. Given the complexity of recognizing actions and activities from images and videos, achieving high accuracy in challenging environments has been a focal point of these efforts. However, there is still a need to develop efficient techniques for human action recognition from images and videos.

In traditional approaches to human activity recognition, machine learning models heavily rely on hand-crafted features. However, this process is challenging and demands a significant level of domain expertise and feature engineering. The advent of deep neural networks has revolutionized this field by enabling models to automatically learn features from raw sensor data. As a result, classification outcomes have substantially improved. In this paper, we present a novel approach for human activity recognition using ensemble learning of multiple convolutional neural network (CNN) models. Two different CNN models are trained on the publicly available datasets. The performance of the proposed model is better than the approaches available in literature review.

The Paper is organized as follows: the related work is explained in section II. The proposed model is explained in section III. The results and discussion about the model is presented in section IV.

2. Review of Literature

Deep learning is a specialized branch of machine learning that employs neural networks to autonomously extract valuable features directly from the data. It has demonstrated impressive achievements in various computer vision tasks,

notably action recognition. Convolutional Neural Networks (CNNs) [2] and Recurrent Neural Networks (RNNs) [3] are frequently employed deep learning models in the field of action recognition. These models excel at learning highly resilient and distinguishing features, surpassing the capabilities of manually crafted features. However, their training process demands substantial volumes of annotated data and significant computational resources.

Numerous methods utilizing deep learning for human action recognition have been extensively explored [4, 5]. In comparison to traditional machine learning approaches for recognizing human behavior, deep learning methods do not rely on specific human experience or knowledge. Instead, they directly identify human actions in videos using end-to-end approaches [6]. These methods can be categorized into two groups based on feature extraction techniques: human action recognition based on skeletons and human action recognition based on feature maps. Notably, spatiotemporal networks and Two-Stream networks are prominent deep learning approaches [7]. Among these approaches, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are widely adopted [8, 9].

Hritam Basak, *et al.* [10] proposed a framework that employs deep learning and swarm intelligence-based meta-heuristic for human action recognition that uses 3D skeleton data for action classification. The model extracts four different types of features from the skeletal data and uses a modified version of Inception-ResNet for classification. The model has been evaluated on three publicly available HAR datasets, achieved competitive results. Zehua Sun, *et al.* [11] presented a comprehensive survey of recent progress in deep learning methods for HAR based on the type of input data modality. It reviews the current mainstream deep learning methods for single data modalities and multiple data modalities, including the fusion-based and the co-learning-based frameworks.

Hao Yang, *et al.* [12] developed a novel network, named Feedback Graph Convolutional Network (FGCN) that introduces a feedback mechanism into GCNs for action recognition. The model provides predictions on-the-fly and has achieved the state-of-the-art performance on three datasets. In [13] Yanan Liu, et al proposed a kernel attention adaptive graph transformer network (KA-AGTN), which models the higher-order spatial dependencies between joints by the graph transformer operator based on multihead self-attention. The model outperforms the baseline 2s-AGCN and achieves the state-of-the-art performance on the Kinetics-Skeleton 400 dataset.

Zehra, Narjis *et al.* [14] presented three CNN based models as well as their ensembles for WISDM dataset of HAR. It was proved that the performance of the ensemble model is better than that of individual models. In [15], authors proposed a deep neural architecture called LSTM-RNN. The network demonstrates the capability to successfully and efficiently learn all six activity classes with a relatively low number of training epochs, achieving high levels of accuracy. The model has been successfully exported and integrated into an Android app, utilizing the Tensorflow library. This implementation allows for real-time predictions to be made. The multiclass SVM approach mentioned in [16] gave an overall accuracy of 94.12%.

In [17], the Resnet-50 Pre-Trained Model is used for extracting features and classification. The features extracted are fused by the Canonical Correlation Analysis (CCA). Then features are selected using the Shanon Entropy-based threshold function. The selected features are finally passed to multiple classifiers for final classification. Experiments are conducted on five publicly available datasets as IXMAS, UCF Sports, YouTube, UT-Interaction, and KTH. They have achieved very good results for these datasets.

3. Proposed Work

The primary objective of this paper is to develop a system for human action recognition using video data. The videos are first segmented, and the individual frames are imported as input data. Two deep learning techniques are employed to generate feature maps specifically designed for human action recognition. During the training phase, the network learns to recognize different human actions and assigns class tags accordingly. To evaluate the performance of the proposed approach, standard datasets such as UCF101 and

HMDB51 are used. In this section, we provide an overview of these datasets, highlighting their characteristics and relevance to the task at hand. Subsequently, we delve into the implementation details of our approach, outlining the specific methodologies employed to achieve accurate human action recognition.

A. Datasets

UCF101 [18] is a comprehensive dataset for action recognition, containing a wide range of realistic action videos sourced from YouTube. It expands upon the UCF50 dataset by including 101 action categories, making it the largest and most diverse collection of its kind. With a total of 13,320 videos, UCF101 offers a remarkable variety of actions, encompassing various factors such as camera motion, object appearance, pose, scale, viewpoint, background clutter, and illumination conditions. This diversity makes UCF101 a highly challenging dataset, pushing the boundaries of action recognition research. Unlike many existing datasets that feature staged performances by actors, UCF101 aims to foster new avenues of study by introducing novel and realistic action categories. The videos in UCF101 are organized into 25 groups, each consisting of 4-7 videos representing a specific action. Videos within the same group often share common characteristics, such as similar backgrounds or viewpoints, providing additional contextual information for analysis. The action categories in UCF101 can be classified into five main types: Human-Object Interaction: Actions that involve interactions between humans and objects, such as "eating," "drinking," or "using a computer." Body-Motion Only: Actions that primarily focus on human body movements without explicit object interactions, including actions like "walking," "jumping," or "running." Human-Human Interaction: Actions involving interactions between two or more humans, such as "hugging," "shaking hands," or "fighting." Playing Musical Instruments: Actions that revolve around playing musical instruments, covering a range of activities like "playing the guitar," "drumming," or "singing." Sports: Actions related to various sports activities, encompassing actions like "basketball-dunk," "soccer-juggling," or "tennis-swing." By providing a diverse and realistic dataset, UCF101 encourages researchers to explore and advance the field of action recognition, enabling the development of more robust and accurate models for real-world applications.

The HMDB51 [19] dataset comprises 51 distinct action categories, with a minimum of 101 clips per category, resulting in a total of 6,766 video clips sourced from a diverse range of origins. It stands as the largest and most realistic dataset available to date, providing valuable resources for action recognition research. Each video clip has undergone validation by a minimum of two human observers, ensuring consistency in the dataset. Furthermore, the HMDB51 dataset includes additional meta information that enhances its utility. This meta information allows for precise selection of testing data and facilitates the training and evaluation of recognition systems. The provided meta tags offer insights into various aspects of the clips, including camera viewpoint, the presence or absence of camera motion, video quality, and the number of actors involved in the action. These meta tags enable researchers to conduct more flexible experiments by utilizing specific subsets of the dataset for evaluating the performance of computer vision systems. Overall, the HMDB51 dataset is a comprehensive collection of action categories, combined with the meticulous validation process and rich meta information, makes it an invaluable resource for advancing the field of computer vision and action recognition. Researchers can leverage this dataset to develop and evaluate more robust recognition systems that can accurately analyze and understand actions in real-world scenarios.

Table 1: Challenges present in UCF101 and HMDB51 Datasets.

Dataset	Background	No. Of Cameras	Camera Movement	View Type	Occlusion	Acted
UCF101	Dynamic	Unspecified	Dynamic	Single	Yes	No
HMDB51	Cluttered	Unspecified	Non-Static	Single	No	Yes

Figure 1 (A) Sample frames of action classes of HMDB51 dataset and (B) Sample frames of action classes of UCF101 dataset.

B. ResNet50+ Custom CNN

The key innovation in the implemented model is the use of ensemble leaning i.e. ResNet50 and a custom convolutional neural network (CNN). Initially the elaborate pre-processing approach is used as it leverages two models for feature extraction. The frames are resized to a uniform dimension before being fed into the models, thereby maintaining data consistency. Every twenty-fifth frame is captured for analysis to ensure temporal relevance while reducing computational costs. This "downsampling" technique is particularly useful when dealing with high frame rate videos where consecutive frames may carry redundant information. It also plays a significant role in making the model more computationally efficient without compromising the quality of the output significantly.

Each model might learn different aspects of the video data, resulting in a more comprehensive representation of the video's content. This could potentially improve the performance of the final action recognition model. By combining the features and training a new model on the top, the idea is to build a model that can capture the strengths of both models and leads to better overall performance.

A Convolutional Neural Network (CNN) is an advanced deep learning algorithm utilized for computer vision tasks, particularly in the recognition and classification of image features. It is specifically designed as a multi-layer neural network to analyse visual inputs, enabling tasks such as image classification, segmentation, and object detection. The CNN comprises two essential components: a Convolutional layer that divides the image into different features for in-depth analysis, and a fully connected layer that utilizes the convolutional layer's output to generate the most accurate description of the image. The structure of a basic convolutional neural network for image classification is illustrated in Figure 2.

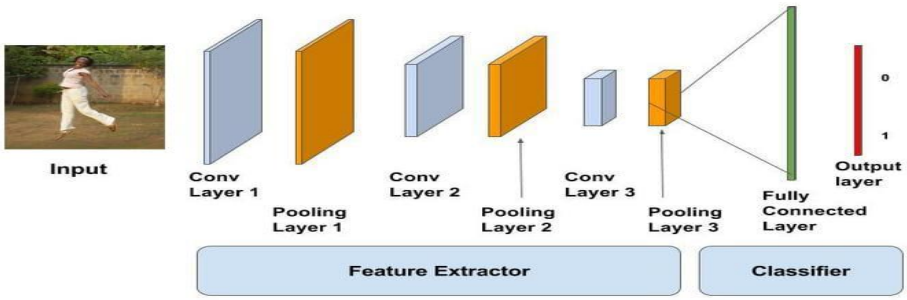

Figure 2. Basic structure of Convolutional Neural Network for image classification.

To address the issue of vanishing gradients that arises when increasing the depth of a neural network, the Residual Network [20] was introduced. This architecture incorporates a clever mechanism where any given layer K not only receives input from the previous layer but also takes in the output of an earlier layer, which is then connected as input to the *Kth* layer. As a result, the output for *the Kth* layer can be expressed as:

$$H(x)l=f(x)+x \quad \text{-- (1)}$$

Here, f(x) represents the function of the lth layer, and x denotes the previous input to the (l-1)th layer. The ResNet architecture comes in several versions, including ResNet34, ResNet50, ResNet101, and ResNet152. ResNet50 is created by replacing each two-layer block in 34 layer 34-layer net with 3-layer bottleneck block. This model has 3.8 billion FLOPs.

The proposed ensemble model's architecture is a testament to creative innovation, with the outputs from two networks being concatenated together. This concatenated layer is then passed through a series of dense, batch normalization, and dropout layers, which serves a dual purpose. The dense layers enhance the model's learning capacity while the batch normalization layers ensure a smooth and speedy training process by controlling the input mean and variance to subsequent layers. Dropout layers are introduced to encourage model robustness by reducing overfitting. Figure 3 represents the structure of proposed model.

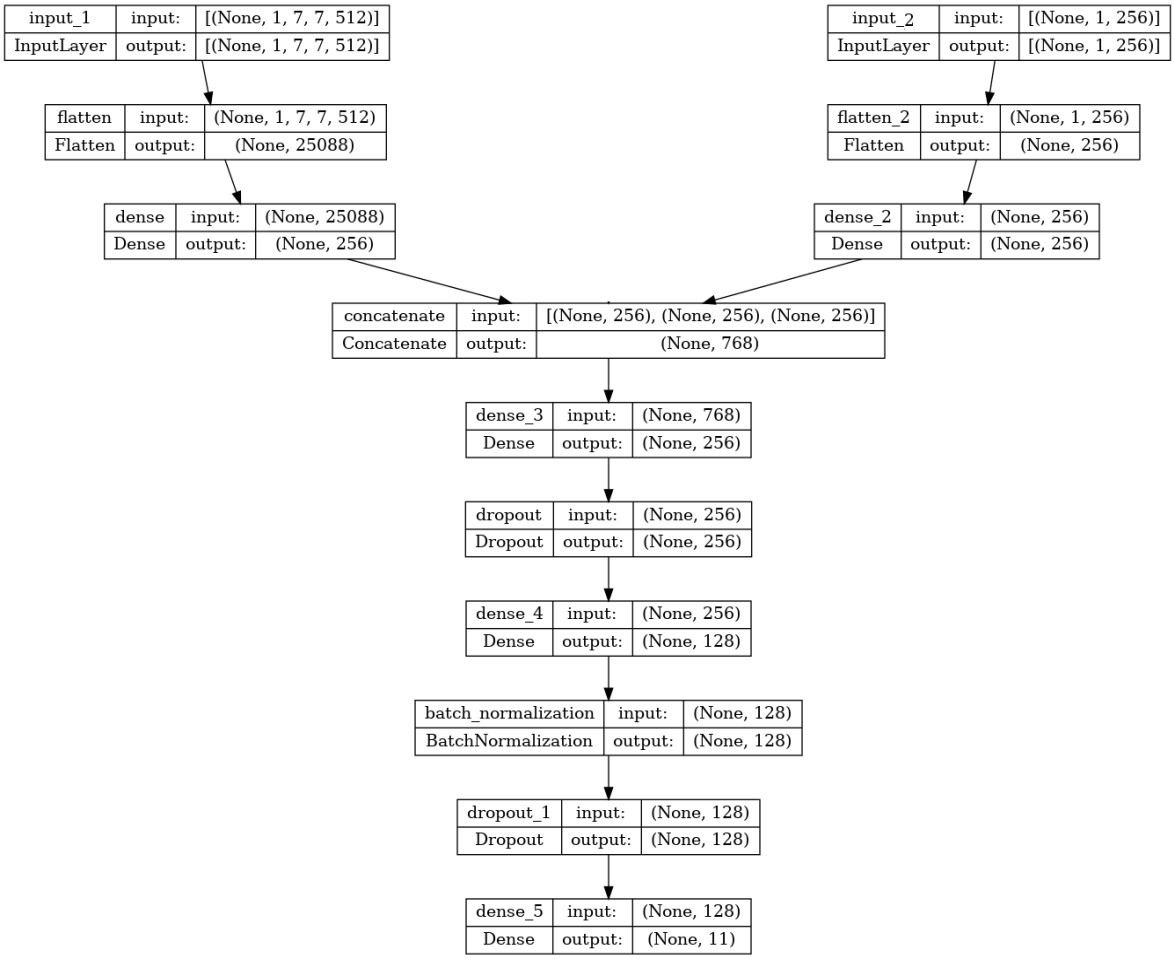

Figure 3. Basic structure of the proposed model.

4. Result and Discussion

In this work, we focus on the application of ensemble learning within the field of deep learning for video action recognition. Our approach is novel due to several distinct and innovative aspects, primarily in the model's architecture, the feature extraction process, and the pre-processing techniques utilized. Traditionally, video action

recognition models primarily rely on a single type of pre-trained model for feature extraction. In contrast, our approach leverages the power of ensemble learning by utilizing a combination of two distinct models: ResNet50, and a custom-designed Convolutional Neural Network (CNN). Each model is uniquely capable of extracting different types of features from input data. ResNet50, with its residual connections, excels in learning high-level features even from deeper layers of the network, ensuring no information is lost through the depth of the model. Our custom CNN, designed with multiple convolutional, pooling, and dropout layers, further boosts the model's capacity to learn more complex and intricate features from the data. This amalgamation of models, each with its unique capabilities, ensures a comprehensive extraction of features from the data, thus, improving the model's predictive performance.

Our proposed model is demonstrating significant improvements over existing models due to its efficient use of resources and superior feature extraction capabilities. The blend of different types of pre-trained and custom model architecture allow it to effectively capture both spatial and temporal features with high precision, making it highly suitable for video action recognition tasks.

Table 2. Comparison with state-of –the- art methods

Authors	**Year**	**Dataset**	**Accuracy**
Sadia Kiran et al.	2021	UT-Interaction	96.7%
Zeqi Yu and Wei Qi Yan et al.	2021	HMDB51	89.74%
Yixue Lin et al.	2020	PASCAL VOC	92.1%
Hritam Basak et al.	2022	UTD-MHAD	97.56%
Hao Yang et al.	2020	Northwestern-UCLA	95.3%
Proposed Model	**2023**	**HMDB51, UCF101**	**99.10%, 98.64%**

Comparing state-of- the-art models with our proposed model, it's clear that our model's unique combination of ensemble learning, feature extraction, and preprocessing techniques, as well as its high accuracy of 99.10%, sets it apart. However, a direct comparison in terms of performance would require a thorough evaluation on the same datasets.

Conclusion

This paper presented a novel and robust system for human action recognition using video data. This was accomplished through the innovative application of ensemble learning that leverages a combination of two deep learning models, ResNet50 and a custom Convolutional Neural Network (CNN). Through the strategic application of these models, the system effectively managed to extract comprehensive features from the video data, thus improving its predictive performance. The approach was tested using two standard and challenging datasets, UCF101 and HMDB51, providing a rigorous evaluation of the system's performance. Results revealed a significant improvement in the recognition accuracy over existing models, which can be attributed to the system's efficient use of resources and superior feature extraction capabilities. In particular, the system's capacity to capture both spatial and temporal features with high precision made it highly suitable for the task of video action recognition.

A comparison with other state-of-the-art models further attested to the proposed model's superior performance, showing an impressive accuracy of 99.10% and 98.64% on HMDB51 and UCF101 datasets, respectively. While this paper provided a strong foundation, further evaluation using a variety of datasets would be beneficial to ensure the system's broader applicability. The results of this work suggest promising future directions in the domain of video action recognition, emphasizing the potential of ensemble learning strategies in improving model performance.

Reference

[1] M. Selmi, M. A. El-Yacoubi, and B. Dorizzi, “Two-layer Discriminative Model for Human Activity Recognition,” IET Computer Vision,vol. 10, no. 4, pp. 273-278, 2016.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol.*86,* no. 11*, pp. 2278-2324, 1998.*

[3] R. J. Williams, and D. Zipser, "Gradient-based learning algorithms for recurrent. Backpropagation: Theory, architectures, and applications*"*, vol. 433, no. 17, 1995.

[4] S. G. Pleshkova, A. B. Bekyarski and Z. T. Zahariev, "Based on artificial intelligence and deep learning hand gesture recognition for interaction with mobile robots*, Proceedings of the National Conference with International Participation (ELECTRONICA),* pp. 1-4,2019.

[5] P. Gao, D. Zhao and X. Chen, "Multi-dimensional data modelling of video image action recognition and motion capture in deep learning framework," IET Image Processing, vol. 14, no.7, pp. 1257-1264, 2020.

[6] B. Zhang, C. Quan and F. Ren, "Study on CNN in the recognition of emotion in audio and images, " *Proceedings of the IEEE/ACIS International Conference on Computer and Information Science (ICIS),* Okayama, pp. 1-5, 2016.

[7] S. Deep and X. Zheng, "Leveraging CNN and transfer learning for vision-based human activity recognition," *Proceedings of the International Telecommunication Networks and Applications Conference,* Auckland, New Zealand, pp.1-4, 2019.

[8] Y. Liu, P. Wang and H. Wang, Target tracking algorithm based on deep learning and multi-video monitoring, *Proceedings of the International Conference on Systems and Informatics (ICSAI)*, pp. 440-444, 2018.

[9] A. Baisware, B. Sayankar and S. Hood, Review on recent advances in human action recognition in video data, *Proceedings of the International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-19)*, pp. 1-5, 2019.

[10] H. Basak, R. Kundu, P.K. Singh, M.F. Ijaz, M. Woźniak, and R. Sarkar, A union of deep learning and swarm-based optimization for 3D human action recognition. *Scientific Reports*, *12*(1), pp.5494, 2022.

[11] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *Proceedings of the IEEE transactions on pattern analysis and machine intelligence,* 2022.

[12] H. Yang, D. Yan, L. Zhang, Y. Sun, D. Li, and S.J. Maybank,. "Feedback graph convolutional network for skeleton-based action recognition," Proceedings of the *IEEE Transactions on Image Processing*, vol. *31*, pp.164-175, 2021.

[13] Y. Liu, H. Zhang, D. Xu, and K. He, "Graph transformer network with temporal kernel attention for skeleton-based action recognition," Knowledge-Based Systems, vol. *240*, p.108146, 2022.

[14] Zehra, Narjis, S. H. Azeem, and M. Farhan, "Human activity recognition through ensemble learning of multiple convolutional neural networks," *Proceedings of the 2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1-5, 2021.

[15] S. W. Pienaar and R. Malekian, "Human activity recognition using LSTM-RNN deep neural network architecture," *Proceedings of the IEEE 2nd Wireless Africa Conference (WAC),* pp. 1–5, 2019.

[16] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," Future Generation Computer Systems, vol. 81, pp. 307–313, 2018.

[17] Kiran, Sadia, M. A. Khan, M. Y. Javed, M. Alhaisoni, U. Tariq, Y. Nam, R. Damaševičius, and M. Sharif, "Multi-Layered Deep Learning Features Fusion for Human Action Recognition," Computers, Materials & Continua, vol. 69, no. 3, 2021.

[18] K. Soomro, A. R. Zamir, and M. Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild." CRCV-TR-12-01, 2012.

[19] Jhuang, Hueihan, G. Juergen, S. Zuffi, C. Schmid, and M. J. Black. "Towards understanding action recognition, " *Proceedings of the IEEE international conference on computer vision*, pp. 3192-3199, 2013.

[20] He, Kaiming, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition," the *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.