RESEARCH ARTICLE

SAMRIDDHI Volume 14, Special Issue 2, 2022 Print ISSN: 2229-7111 Online ISSN: 2454-5767

Human Activity Recognition Using Feature Fusion

Rashmika K. Vaghela[1*], Jigar A. Patel[2] and Kirit Modi[3]

[1,2]Department of Computer Science and Engineering, Parul University, Vadodara, India

[3]Department of Computer Science and Engineering, Sankalchand Patel College of Engineering, India

Abstract

Human endeavour Since they are and will be used in many new innovations, not only for security and surveillance but also for understanding human behavioural patterns, recognition is becoming more crucial in today's society. The goal of this research project is to create an intelligent model that can recognise human activity from video input from various sources, such as CCTV camera recorded video or YouTube video. Several techniques for identifying human activity have been described in recent years employing sensor-based datasets, depth, skeleton, and RGB (red, green, and blue) datasets. The majority of approaches for classifying activities using sensor-based and skeletal datasets have limitations in terms of feature representation, complexity, and performance. The provision of an effective and economical approach for human activity recognition utilising a video dataset, however, remains a difficult topic. In this research, we propose a frame processing derived from video files for action discrimination by capturing geographical information and temporal changes. To extract discriminative features, we perform transfer learning using pretrained models (VGG19 and DenseNet121 trained on the ImageNet dataset), and we assess the suggested approach using a number of fusion techniques. Using the UCF-50 dataset, our deep learning-based approach is effective. We achieve accuracy of between 95% and 98%.

Keywords: Recognition of human actions/activities (HAR), deep learning techniques, Convolutional Neural Networks, VGG19, DenseNet121

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2022); DOI: 10.18090/samriddhi.v14spli02.25

Introduction

The study of computer vision and pattern recognition has created an interest in the area of video-based human action recognition.[1] Many different fields, including surveillance, robotics, healthcare, video searching, and human-computer interaction, are among its many potential uses. Human action detection in videos faces several difficulties, including crowded backdrops, occlusions, viewpoint fluctuation, execution rate, and camera motion. Over the years, numerous strategies have been put up to deal with the difficulties. Three different dataset types are used for research-single viewpoint, multiple viewpoints, and RGB-depth videos. This paper provides an overview of several cutting-edge deep learning-based methods for the recognition of human actions on the three different kinds of datasets. This review provides information on current trends and potential paths for future work to aid academics in light of the rising popularity and recent advancements in video-based human action identification.[2] Human activity recognition (HAR) is one of the most challenging problems in computer vision. Determine the actions and activities of the person using an intelligent video system. The human-computer interface, tracking, security, and health monitoring are only a few applications for this action monitoring system.

Corresponding Author: Rashmika K. Vaghela, Department of Computer Science and Engineering, Parul University, Vadodara, India, e-mail: rashmivaghela.rv@gmail.com

How to cite this article: Vaghela, R.K., Patel, J.A. and Modi, K. (2022). Human Activity Recognition Using Feature Fusion. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* Volume 14 Special Issue (2), 288-293.

Source of support: Nil

Conflict of interest: None

Despite increasing improvements in this field, detecting activity in an uncircumscribed territory is still difficult with many difficulties. In order to identify various human activities, recent research papers with various methodologies are analyzed throughout this article. This endeavour requires the use of mobile phone sensors and wearable technology. The researchers say the vision-based approach has become a common HAR strategy.[3] A fundamental deep learning network is CNN. Human action recognition from digital photos or video frames is used to identify a tale in a single frame. By applying the end-to-end method, CNNs are used to classify a series of moving images into a category of human activities.

Figure 1: Sample image frames of Activity performed by human from three different class

Related Work

This section describes various methods used for human activity recognition by various researchers:

Due to developments in information and communications technology, the Internet of Things (IoT) has gained considerable popularity and completely changed the field of study known as Human Activity Recognition (HAR).[4] Vision-based and sensor-based approaches can present superior data for the HAR job, but at the expense of user annoyance and social restrictions like privacy concerns. The use of WiFi in intelligent daily activity monitoring for elderly people has grown in popularity in contemporary healthcare applications due to the widespread availability of WiFi devices. One of the properties of WiFi signals is Channel State Information (CSI), which can be used to identify various human activities. In order to collect CSI data for seven distinct human everyday activities, they used a Raspberry Pi 4. They then transformed the CSI data into images and used the resulting images as the inputs for a 2D Convolutional Neural Network (CNN) classifier. Testing has demonstrated that the suggested CSI-based HAR outperforms rival techniques such as 1D-CNN, Long Short-Term Memory (LSTM), and Bi-directional LSTM, achieving an accuracy of about 95% for seven activities.

This research suggests a novel deep neural network for recognizing human behavior that blends LSTM and convolutional layers.[5] The fully-connected layer is primarily the focus of the CNN weight parameters. IN RESPONSE TO THIS CHARACTERISTIC, a GAP layer is employed in place of the fully-connected layer behind the convolutional layer, drastically reducing the model parameters while maintaining a high recognition rate. A BN layer is added after the GAP layer, and an obvious effect is obtained to speed up the model's convergence. The raw data gathered by mobile sensors is input into a two-layer LSTM followed by convolutional layers in the suggested architecture, enabling it to learn the temporal dynamics on different time scales in accordance with the learnt LSTM parameters for improved accuracy. The three open datasets, UC-HAR, WISDM, and OPPORTUNITY, were utilised in the experiment to demonstrate the proposed model's ability to generalize and efficacy. The F1 score was employed to assess the model performance because accuracy is not a suitable and comprehensive measure of performance. Finally, on the UCI-HAR, WISDM, and OPPORTUNITY datasets, the F1 score was 95.78%, 95.85%, and 92.63%, respectively. They also investigated the effects of several hyper-parameters on model performance, including batch size, kind of optimizer, and number of filters. Finally, the model was trained using the best hyper-parameters for the final design. In conclusion, the LSTM-CNN model consistently outperforms those suggested in other research and exhibits sound generalization. Under the assumption of a few model parameters, it can avoid difficult feature extraction and has good recognition accuracy.

This research proposes a novel method for recognizing human actions by combining deep learning with spatiotemporal picture creation from 3D skeletal joints.[6] They examine the joints of the 3D skeleton and suggest mapping the line between the identical joints in two adjacent frames to encode the spatiotemporal image from the joints. The spatial and temporal information is extracted by maintaining the geometry of the movement and joining the line with various colours along with the temporal changes. They use pre-trained deep learning models to assess the value of the proposed method's spatiotemporal representation. The studies are carried out using two different approaches: (i) individual views (front, side, and top views); and (ii) fusion methods (average, multiplication, and maximization). According to the experimental findings using individual views, the front view dataset appears to function well. When using feature fusion, maximization greatly raises the recognition rate. To demonstrate the durability of our work, they also compare the recognition accuracy with three deep learning models. Views and the pace of the action have no effect on the features extracted from the spatiotemporal image. In light of this, both the pre-trained light-weight and heavy-weight deep learning models may easily individualize the actions. The suggested method outperforms cutting-edge efforts using the UTD-MHAD and MSR-Action3D benchmark skeleton datasets, despite the fact that they conduct experiments with individual views along the XY, YZ, and ZX planes. The experimental findings are also investigated using three distinct fusion approaches to determine the best strategy. The suggested approach performs better overall in pre-trained deep-learning experiments on the UTD-MHAD and MSR-Action3D skeletal datasets.

For six activities of daily living using the WISDM dataset, they have presented a CNN model and an LSTM model in this study with accuracy rates of 99.593 and 84.71%, respectively. [7] These two models are quick and accurate thanks to the use of Conv2D layers for CNN, Dropout regularisation, and perfect model hyperparameters in their networks. The authors suggest implementing this Human Activity Recognition framework as a solution for a smart IoT-based monitoring system for eldercare or childcare as future works. Additionally, it will be a fantastic assignment if they can create our own dataset for a predetermined number of common activities individuals engage in on a daily basis using the proper sensors and software. This research topic appears to have numerous cutting-edge Deep Learning applications

in the near future. Additionally, authors propose using the reinforcement learning paradigm to classify and recognize activities in future works.

For Human Activity recognition not much research work done on video dataset till now. Most of the work is done on sensor-based, human pose, and image datasets. In case wearable sensor-based approach can be considered as a good solution for HAR. However, it faces a problem of low recognition rate, managing the vast number of information that the devices can produce, as well as their temporal dependency, and, second, the lack of knowledge about how to relate this data to the defined movements also the wrong placement or orientation of sensors could be causing a problem or effect the recognition performance.

From this study we found that pose dataset approach gives better results in such scenarios but there is no appropriate pose dataset available on the internet for training purposes and classification activities.

We can use image dataset for human activity classification purpose but still single image in not enough for classifying the maximum activity performed by human. Pre-activity and post-activity information are also necessary(temporal) for correct activity recognition, to achieve better results with more computational power then a video processing -based approach will become the best solution for human activity recognition Figure 1 represents sample image frames which can be used for human activity classification from three different activity classes.

Proposed System

In this section, we describe the proposed human activity recognition system based on pre-trained models and Transfer Learning. Working of our proposed model is as shown in Figure 2.

In this research, we are using UCF50 dataset for human activity recognition. UCF 50 dataset contains:50 Action Categories consisting of realistic YouTube videos.

Video preprocessing

First, we extract the videos from the selected classes and create the required dataset.

Like features- A list containing the extracted frames of the videos which are resized and normalize according to our model requirements. Labels: A list containing the indexes of the classes associated with the videos. Video_files_paths: A list containing the paths of the videos in the disk.

Backbone Models

VGG19

VGG19 is a deep convolutional neural network architecture (as shown in Figure 3) designed for image classification and object recognition tasks. It is part of the VGG (Visual Geometry Group) family of architectures developed by researchers at the University of Oxford. VGG19 is an extension of the original VGG16 architecture, with 19 layers, including convolutional and fully connected layers.

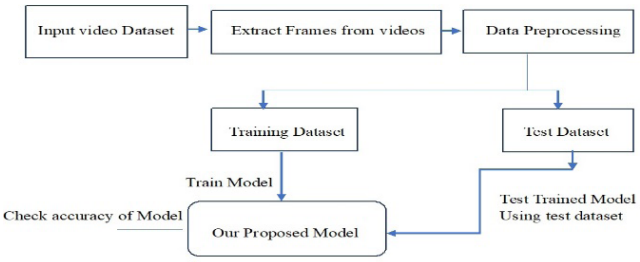

Figure 2: Working of Our Proposed Model

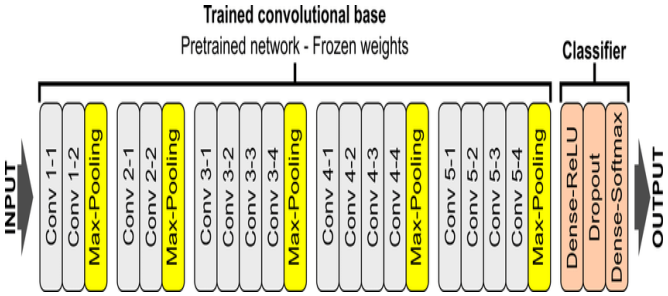

Figure 3: VGG19 Architecture

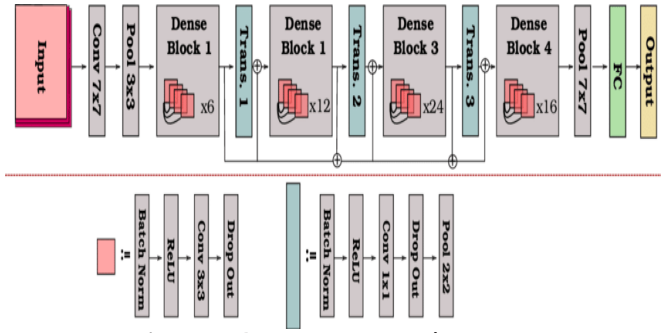

Figure 4: DenseNet121 Architecture

DenseNet121

DenseNet (Densely Connected Convolutional Networks) is a deep neural network architecture (as shown in Figure 4) that introduces the concept of dense connections between layers. DenseNet architectures are designed to address some of the challenges of gradient vanishing and feature reuse that can occur in traditional convolutional neural networks (CNNs). DenseNet121 is one variant of the DenseNet architecture.

DenseNet121 specifically refers to a variant of the DenseNet architecture with approximately 121 layers. It has been pre-trained on large image datasets such as ImageNet and can be fine-tuned or used for transfer learning on a wide range of computer vision tasks. DenseNet architectures are known for their strong performance, efficient use of parameters, and ability to capture rich feature representations, making them popular choices for various deep learning applications.

Transfer learning

Fine-tuning is a transfer learning technique commonly used in deep learning to improve the performance of pre-trained models on a specific task. It involves adjusting the weights

of some of the layers in the pre-trained model while keeping the weights of other layers frozen. This allows the model to learn task-specific features while retaining the knowledge captured by the pre-trained model.

Implementation And Results

The experiment conducted in this research used Kaggle.com which is online deep learning platform. In this we run our system on GPU P100 provided by portal for better and fast execution.

Steps in Our proposed model

- Import the required modules and libraries, including the Keras parts required to create and train the model.
- Define constants, such as the number of classes (NUM_CLASSES) and the sequence length (SEQUENCE_LENGTH).
- Load models from DenseNet121 and VGG19 (without top layers) that have already been trained on ImageNet. The foundation models for feature extraction will be these models.
- Freeze all base model layers to stop them from changing during training.
- To process each frame in the input sequence independently, apply the TimeDistributed wrapper to the underlying models.
- To produce a feature vector for each frame, apply GlobalAveragePooling2D over the sequence dimension for each model.
- Add fully linked (Dense) layers with ReLU activation to the feature vectors for compatibility.
- Adding or multiplying the feature vectors from the two basis models to get an amalgamated feature representation.

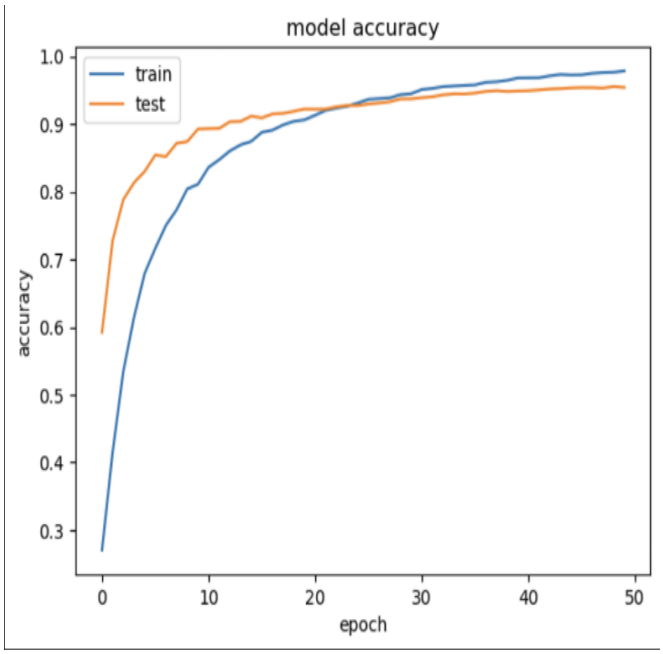

Figure 5: Accuracy Chart for Merge Feature Fusion

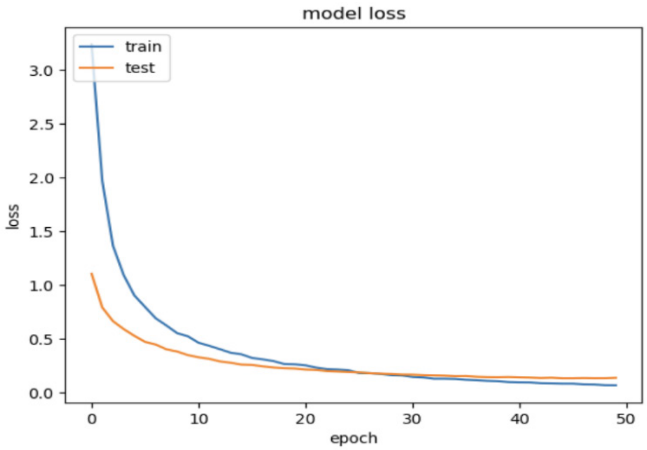

Figure 6: Loss Chart for Merge Feature Fusion

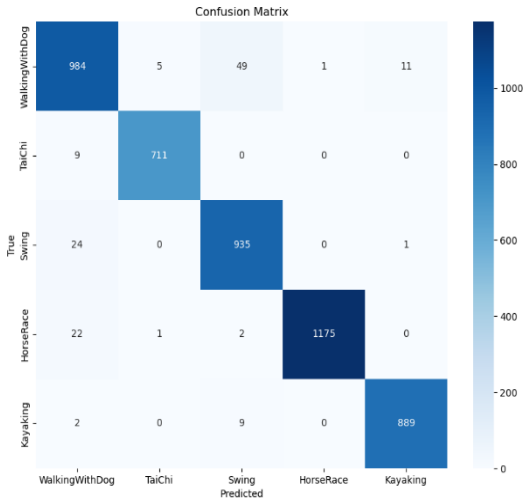

Figure 7: Confusion Matrix for Merge Feature fusion

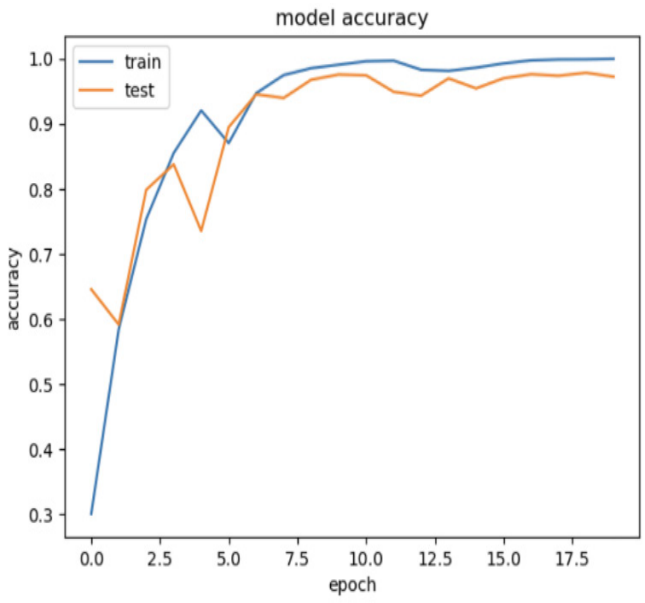

Figure 8: Accuracy Chart for Multiply Feature Fusion

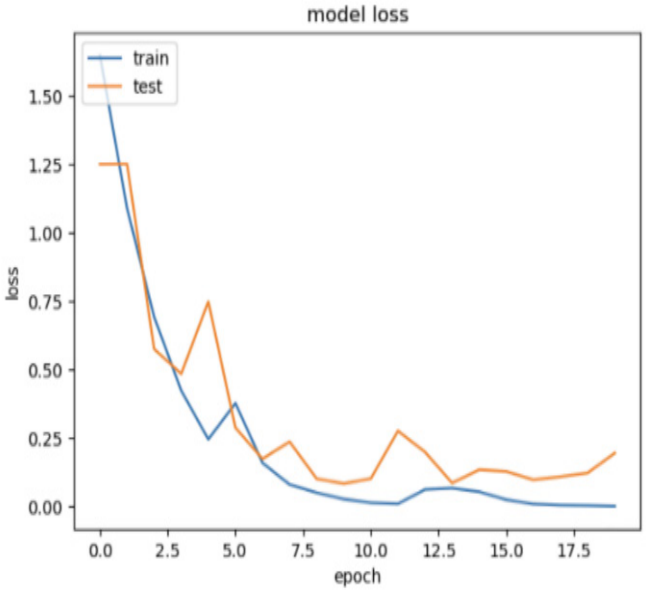

Figure 9: Loss Chart for Multiply Feature Fusion

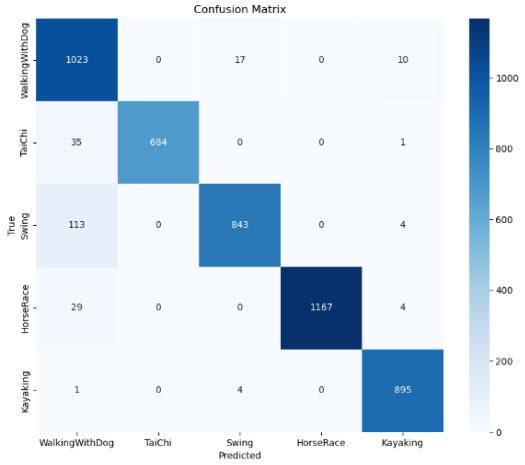

Figure 10: Confusion Matrix for Multiply fusion

Table 1: Results related to work done:

Paper	*Used Method*	*Dataset*	*Accuracy For HAR System*
LSTM-CNN Architecture for Human Activity Recognition (2020)[5]	CNN + LSTM	UCI-HAR	95.78%
Human Activity Recognition Using CNN & LSTM (2020)[7]	CNN	WISDM	99.59
	LSTM		84.71
Our Model	DenseNet121 + VGG19 With Feature fusion	UCF50	95 to 98 %

- Add additional Dense layers for fine-tuning and Dropout layers for regularisation.
- Make the combined model using the inputs and outputs that were given.
- The Adam optimizer and categorical cross-entropy loss are used to compile the model.
- Print out a description of the model architecture.
- By making certain layers of both base models too trainable, you may fine-tune them.
- After fine-tuning, use an optimizer with a reduced learning rate to recompile the model.
- Create an Early Stopping callback to keep track of validation loss and restore the best weights.
- Using the training data, train the combined model with fine-tuning.
- Plot the training history to see model accuracy and loss during training. Evaluate the fine-tuned model on the test data to determine loss and accuracy.

After implementation of both feature fusion methods, we got 95% to 98% accuracy for human activity classification as depicted in Figure 5 and Figure 8. We have also shown loss graph for both the methods in Figure 6 and Figure 9. We have implemented our proposed methods for 05 classes of human activity. Based on Confusion Matrix showed in Figure 7 and Figure 10 , we calculated Precision, Recall and F1-score for each activity class.

Conclusion

The proliferation and adding of features with optimized parameter settings produce good results with video files. 95 to 98% accuracy has been attained.

In conclusion, as per Table 1, accuracy results show that the suggested model is the most successful model for identifying human activities based on video inputs. Its robustness in detecting and classifying the diverse human activities included in the dataset is indicated by its high accuracy and low loss. The suggested model functions with video files directly. Variables like processing capabilities, deployment restrictions, and the desired balance between accuracy and efficiency may influence the model of choice.

References

[1] Wu D., Sharma N., and Blumenstein M. (2017). Recent advances in video-based human action recognition using deep learning: A review. International Joint Conference on Neural Networks (IJCNN), pp. 2865-2872.

[2] Banjarey K., Sahu S. P., and Dewangan D. K.(2021). A survey on human activity recognition using sensors and deep learning methods. 5th international conference on computing methodologies and communication (ICCMC), pp. 1610-1617.

[3] Yu Z. and Yan W. Q. (2020). Human action recognition using Deep learning methods.35th International Conference on Image and Vision Computing New Zealand (IVCNZ) pp. 1-6.

[4] Moshiri P. F., Shahbazian Nabati R., M., and Ghorashi S. A. J. S.(2020). A CSI-based human activity recognition using deep learning. vol. 21, no. 21, p. 7225, 2021.

[5] Kun xia , jianguang huang , and hanyu wan.(2020). LSTM-CNN Architecture for Human Activity Recognition. 2020, p. 56855.
[6] Nusrat Tasnim 1 , Mohammad Khairul Islam 2 and Joong-Hwan Baek. (2021). Deep Learning Based Human Activity Recognition Using Spatio-Temporal Image Formation of Skeleton Joints. Appl. Sci. 2021, 11, 2675
[7] Chamani Shiranthika, Nilantha Premakumara, Huei-Ling Chiu, Hooman Samani, Chathurangi Shyalika, Chan-Yun Yang. (2020). Human Activity Recognition Using CNN & LSTM. IE2020 IEEE
[8] Park J., Jang K., and Yang S.-B.(2018). Deep neural networks for activity recognition with multi-sensor data in a smart home. IEEE 4th World Forum on Internet of Things (WF-IoT), 2018, pp. 155-160.
[9] Ramanujam E., Perumal T., and Padmavathi S. J. I. S. J. (2021) . Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. vol. 21, no. 12, pp. 13029-13040.
[10] Singh S. P., Sharma M. K., Lay-Ekuakille A., Gangwar D., and Gupta S. J. (2020). Deep ConvLSTM with self-attention for human activity decoding using wearable sensors. Vol. 21, no. 6, pp. 8575-8582.
[11] Krizhevsky A., Sutskever I., and Hinton G. E. J. A. i. n. i. p. s.(2012). Imagenet classification with deep convolutional neural networks. vol. 25, 2012.
[12] Yang J., Nguyen M. N., San P. P., X. Li, and Krishnaswamy S.(2015). Deep convolutional neural networks on multichannel time series for human activity recognition. Ijcai, Buenos Aires, Argentina 2015, vol. 15, pp. 3995-4001.
[13] Ramanujam E. T., Perumal, and Padmavathi S. J. I. S. J.(2021). Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. vol. 21, no. 12, pp. 13029-13040.
[14] Nweke H. F., Teh Y. W., Al-Garadi M. A., and Alo U. R. J. E. S. w. A. (2018). Deep learning algorithms for human activity recognitionc using mobile and wearable sensor networks: State of the art and research challenges. vol. 105, pp. 233-261, 2018.
[15] Sathyanarayana S., Satzoda R. K., Sathyanarayana S., Thambipillai S. J. J. o. A. I., and Computing H.(2018). Vision-based patient monitoring: a comprehensive review of algorithms and technologies. Vol. 9, pp. 225-251.
[16] Wang J., Chen Y., Hao S., Peng X., and Hu L. J. P. r. l. (2019). Deep learning for sensor-based activity recognition: A survey. Vol. 119, pp. 3-11.
[17] Needell D., Nelson A. A., Saab R., and Salanevich P. J. a. p. a. (2020). Random vector functional link networks for function approximation on manifolds 2020.
[18] Deep S. and Zheng X. (2019). Leveraging CNN and transfer learning for vision-based human activity recognition, 29th International Telecommunication Networks and Applications Conference (ITNAC), 2019, pp. 1-4.
[19] Gao P., Zhao D. and Chen X. J. I. I. P. (2020). Multi-dimensional data modelling of video image action recognition and motion capture in deep learning framework. Vol. 14, no. 7, pp. 1257-1264.
[20] Tufek N., Yalcin M., Altintas M., F., Kalaoglu Y. Li, and Bahadir S. K. J. I. S. J. (2019). Human action recognition using deep learning methods on limited sensory data. Vol. 20, no. 6, pp. 3101-3112.
[21] Chung S., Lim J., Noh K. Kim J., G., and Jeong H. J. S. (2019). Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning. Vol. 19, no. 7, p. 1716.