INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 2 - V8I2-1320)
Available online at: https://www.ijariit.com

Vison-based human activity recognition

M. Madhusudhan
madhusudhan.cse@cmrtc.ac.in
CMR Technical Campus, Hyderabad, Telangana

B. Manish Kumar
177r1a05j5@cmrtc.ac.in
CMR Technical Campus, Hyderabad, Telangana

P Rohit
177r1a0599@cmrtc.ac.in
CMR Technical Campus, Hyderabad, Telangana

V. Sri Ram Reddy
167r1a05p7@cmrtc.ac.in
CMR Technical Campus, Hyderabad, Telangana

V. Sai Chandana
177r1a0559@cmrtc.ac.in
CMR Technical Campus, Hyderabad, Telangana

ABSTRACT

There have been substantial advances in HAR (Human Activity Recognition) in the past several years due to the advancement of the IoT (Internet of Things). HAR may be used in a range of contexts, including elder care, surveillance systems, and anomalous behavior detection. Different machine learning techniques have been used to anticipate human actions in a particular circumstance. Feature engineering techniques, which may pick an ideal collection of features, have outperformed typical machine learning techniques. Deep learning models, like CNN (Convolutional Neural Networks), on the other hand, are known to extract features and minimize computing costs automatically. We employ the CNN model for predicting actions from the Weizmann Dataset in this article. To extract deep image features and trained machine learning classifiers, transfer learning is used in particular. We found that VGG-16 has an accuracy of 96.95 percent in our experiments. We also found that VGG-16 outperformed the rest of the CNN models that were used in our experiments.

Keywords: *Convolutional Neural Network, Activity recognition, Deep Learning.*

1. INTRODUCTION

HAR is a prominent study topicdue to its wider applications in surveillance systems, automated homes, and elderly care. Human activity recognition has been the subject of several topics in the last few years. Existing works are either wearable or non-wearable. A wearable HAR system employs sensors that are fixed on the body. The nature of wearable-based HAR systems is intrusive. Non-wearable HAR does not need any sensors to be attached to the person or the carrying of any device to recognize the activity. These systems are further classified into two types: (i) Sensor-based and (ii) vision-based. The sensor-based system identifies human activity using RF signals from sensors including Wi-Fi signals, PIR sensors, and RFID [1]. To identify human actions, a vision-based system uses images, video frames from IR, or depth cameras. Sensor-based HAR systems are non-intrusive but can't offer a higher level of accuracy. As a result, vision-based systems are gaining popularity at present, however, extracting human activities from the live video is difficult.

According to motion features, video-based activity recognition may be classified as marker-based or vision-based. The optic wearable marker-based Mocap (motion capture) system is utilized in the marker-based technique. However, it can correctly record complicated human actions, this method has certain drawbacks. There must be an optical sensor fixed on the human body with different camera settings. The depth or RGB image is used in the vision-based approach. It does not need the use of external devices or the attachment of sensors to the body of the user. Since this method is gaining prominence, the HAR system is becoming simpler and easier to apply in a variety of settings [2].

There are a few vision-based HAR systems that use typical machine learning approaches for activity detection. The usage of deep learning techniques instead of standard approaches to machine learning has increased significantly. CNNs are often used in computer vision applications. Images are processed using a sequence of convolutional layers [3]. From the Weizmann Dataset, we utilize CNN to identify human activities. The frames for each action were first retrieved from the videos. Transfer learning is used to acquire deep image features as well as trained machine learning classifiers. To categorize activities, we used three different CNN algorithms and compared our findings to previous work on the same dataset.

2. RELATED WORK

Vision-based human activity identification has recently gotten significant attention. Handcrafted feature extraction from videos/images and conventional classifiers for activity identification has been used in the majority of the studies. In many cases, conventional methods yielded the best outcomes and showed the highest levels of performance. Traditional approaches, on the other hand, are impractical to utilize in the real world since handcrafted characteristics are heavily reliant on data and are not flexible enough to adapt to changing conditions [4].

Because of its ability to decode temporal patterns, HMM ("Hidden Markov Model") approaches are widely employed as recognition methods in recent years. However, researchers are increasingly turning to deep learning algorithms because of their proficiency in extracting features automatically and learning deep pattern frameworks. In the computer vision field, deep learning algorithms have ruled out classical categorization techniques. These algorithms have recently significantly gained attention in the computer vision field, with excellent results. Consequently, video-based human activity identification using deep learning algorithms has received considerable interest in recent years.

A mixed-norm regularization function may be added to a deep LSTM network, Zhu et al. suggested an action classification technique. CNNs are one of the most widely used deep learning approaches in frame/image processing. Several studies have used 2D-CNNs, which take advantage of spatial correlation between video frames and integrate the results using several methodologies. Many people have employed optical flow as an extra input to 2D-CNN to gain temporal correlations information [IO]. Subsequently, 3D-CNNs were introduced, and they showed remarkable performance in video and frame classification.

Wang et al. used CNN to extract features from depth and RGB frames automatically. The collected features were fed into a fully connected neural network, which resulted in a higher level of accuracy. Ji et al. suggested a 3D CNN model for activity recognition that performs 3D convolutions and extracts temporal and spatial properties by recording motion data. ConvNet, a two-stream convolution layer design devised by Simonyan et al., may obtain excellent results despite insufficient training data.

Khaire et al. developed an algorithm to detect activities by training convnets using RGB-D datasets and combining SoftMax scores from depth, skeleton, as well as motion images at the classification level. Over a 4D video chunk, Karpathy et al. suggested extending CNN architecture in the first convolutional layers. While Tran et al. employed a deep 3D-CNN model, (like VGG net) to increase the model's accuracy by using spatiotemporal convolutions and pooling in all layers.

In contrast, we're more interested in seeing how transfer learning may be employed with CNN models for improving classification accuracy on a benchmark dataset [5].

3. TRANSFER LEARNING

Transfer learning is a technique for transferring information from previous extensive training to the present model. Transfer learning allows deep network algorithms to be trained with considerably fewer data. It was utilized to shorten the training time and increase the model's accuracy. We employ transfer learning in this paper to use information from large-scale datasets like ImageNet. The frames for each action are first extracted from the videos. Transfer learning is used to acquire deep image features as well as trained machine learning classifiers. The pre-trained weights on ImageNet are utilized as the preliminary step for transfer learning in all CNN models. ImageNet is a database of 2000 images (activities from 1 category). Knowledge is transferred from ImageNet weights to Weizmann datasets since the actions detected in this study fall within the ImageNet domain. The penultimate CNN layer is used to extract the features. Figure 1 illustrates the fundamental concept of transfer learning.

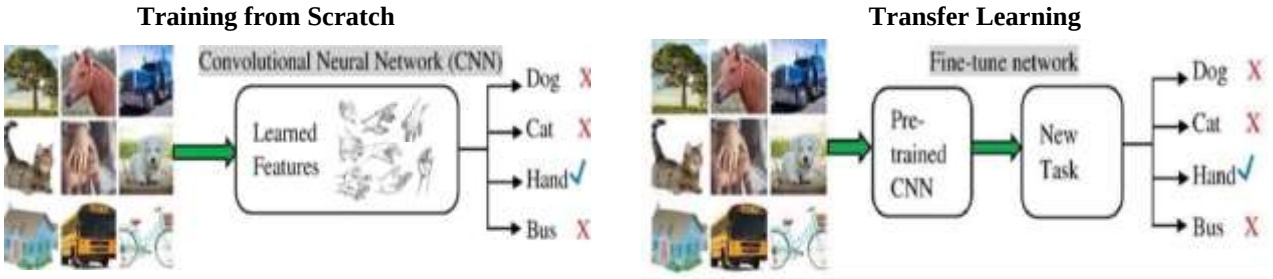

Fig. 1. Schematic representation of transfer learning

In transfer learning, the pre-trained neural model from a large-scale dataset is preserved while the weights are updated in the trained model and the pre-trained neural model is used to extract features.

4. IMPLEMENTATION

A. Dataset

We use trials on the Weizmann dataset to test the models' performance in terms of activity recognition. It consists of 90 low-resolution video frames depicting 9 different individuals doing ten different activities, including bend, jack (or jumping-jack), jump (or jump-forward-on-2-legs), pjump (or jump-in-place-on-2-legs), run, side (or gallop-sideways), slip, well, wave I (wave 1-hand), and wave2 (wave 2-hands). For our experiment, we employed nine different actions (excluding "pjump jump-in-place-on-2-legs"). All videos are first converted into separate frames depending on activity. Table 1 displays the total frames/activity for all 9 participants based on the extracted frames. The complete dataset is subdivided into three sections: testing 20%, training 70%, and validation 10%.

Table-1: Statistics for the dataset in terms of total frames/activity

Activity	No. of frames
Jack	729
Bend	639
Run	346
Jump	538
Skip	378
Side	444
Wave1	653
Wave2	624
Walk	566
Total	4917

B. Discussion and Results

We test three distinct CNNs for activity identification, such as Google's InceptionNet-v3, VGG-16, and VGG-19, to categorize activities. To use the information obtained from large-scale datasets like ImageNet, we applied transfer learning. Using the information gained from pre-trained weights on ImageNet, we experimented on the Weizmann dataset. CNNs' penultimate layers are used to extract the features. We used transfer learning on the VGG-16 CNN algorithm and obtained a 96.95 percent accuracy. VGG-16 takes a 224x224 image as an input and extracts features from the fc1 layer, yielding a 4096-D vector for each image.

To compare the performance of the various CNN models, we employed transfer learning on additional CNN models including Google's InceptionNet-v3 and VGG-19. VGG-19 obtained 96.54 percent and Google's InceptionNet-v3 obtained 95.63 percent, correspondingly. The experimental findings show that VGG-16 outperforms the other CNN models once transfer learning has been applied to all of the models. The accuracy, precision, f1-score, and recall of given models are reported in Table 2. Figures 2, 3, and 4 demonstrate the confusion matrix of three distinct CNN models [6].

On the Weizmann dataset, we were able to match the performance of other systems that did not include transfer learning. When using transfer learning to recognize the same dataset, the experiment results indicated that recognition scores improved. Transfer learning improves recognition accuracy by 1% to 6%. Table 3 compares the performance of the VGG-16 model using transfer learning to those of the other techniques. Transfer learning is compared against state-of-the-art techniques to see how successful it is when used with CNN algorithms for increasing recognition results [7].

Table-2: Activity recognition results based on several CNN models

Model	Precision	Accuracy	Recall	Precision	F1-Score
VGG-16	97.00%	96.95%	97.00%	97.00%	97.00%
Inception-v3	96.00%	95.63%	96.00%	96.00%	96.00%
VGG-19	97.00%	96.54%	97.00%	97.00%	96.00%

Table-3: Comparison of performance using the Weizmann dataset

Model	VGG-16	Kumar et al.	Cai et al.	Han et al.	Feng et al.
Accuracy	96.95	95.69%	95.70%	90.00%	94.10%

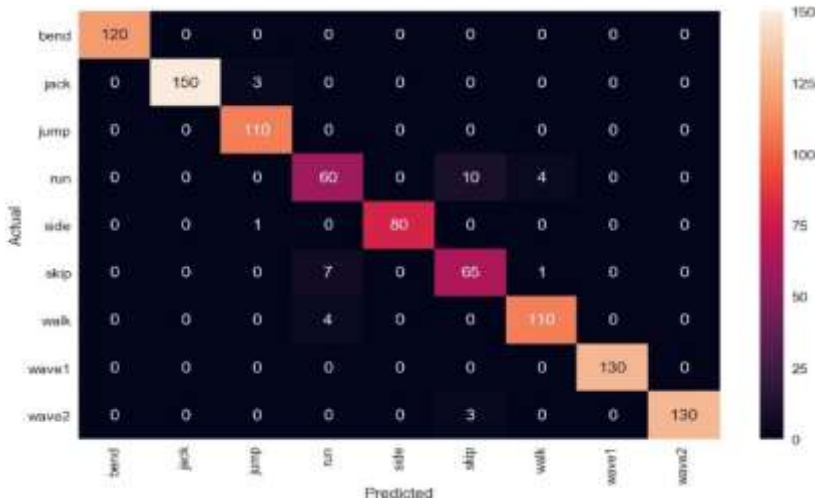

Fig. 2. Confusion Matrix for recognizing 9 activities with VGG-16 CNN on Weizmann Dataset

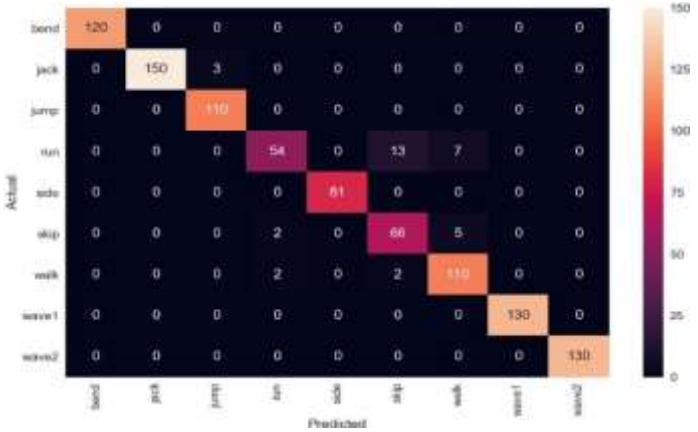

Fig. 3. Confusion Matrix for recognizing 9 activities with VGG-19 CNN on Weizmann Dataset

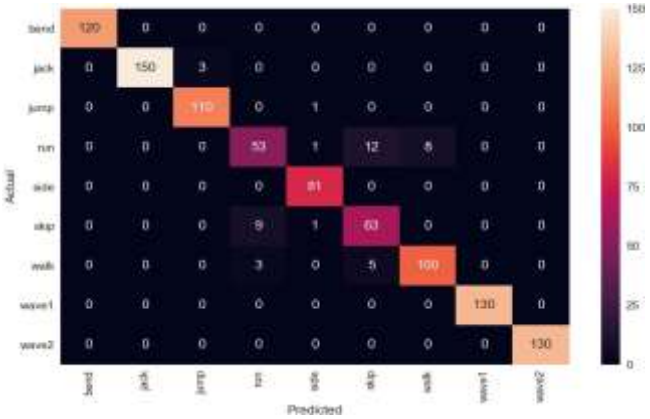

Fig. 4. Confusion Matrix for recognizing 9 activities with Inception-v3 CNN on Weizmann Dataset

Figures 2, 3, and 4 illustrate the confusion matrix of three distinct CNNs after transfer learning, which has been employed to categorize frames of various activities with Google's Inception Net-v3, VGG-16, and VGG-19, correspondingly. Figures 2, 3, and 4 show that VGG-16 has misclassification to forecast running activity as skip, VGG-19 has misclassification to forecast running activity as skip and skip as walk, and Google's InceptionNet-v3 has misclassification to forecast running activity as skip, all of which are visually comparable. The accuracy of activity identification has improved because of the use of transfer learning on CNN models. However, since ImageNet includes photos from different categories, the transfer learning approach employed in our research using information transferred from the pre-trained weight on ImageNet may be compromised.

5. CONCLUSION

This dataset was analyzed using a CNN model to predict human actions. Three CNNs were tested for activity recognition. To extract the deep image features as well as train machine learning classifiers, we have used transfer learning techniques. In our experiments, we found a 96.95 percent accuracy rate utilizing VGG-16 and transfer learning. In our experiments, VGG-16 performed better in comparison to CNN models to extract features. Additionally, we found that VGG-16 outperformed current best practices in our experiments using the transfer learning method.

6. REFERENCES

[1] B. Bhandari, J. Lu, X. Zheng, S. Rajasegarar, and C. Karmakar, "Noninvasive sensor-based automated smoking activity detection," in Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017, pp. 845–848.

[2] L. Yao, Q. Z. Sheng, X. Li, T. Gu, M. Tan, X. Wang, S. Wang, and W. Ruan, "Compressive representation for device-free activity recognition with passive RFID signal strength," IEEE Transactions on Mobile Computing, vol. 17, no. 2, pp. 293–306, 2018.

[3] Lillo, J. C. Niebles, and A. Soto, "Sparse composition of body poses and atomic actions for human activity recognition in RGB-d videos," Image and Vision Computing, vol. 59, pp. 63–75, 2017.

[4] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton-based action recognition using regularized deep LSTM networks," in ThirtiethAAAI Conference on Artificial Intelligence, 2016.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[6] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 2009, pp. 248–255.

[7] S. Deep and X. Zheng, "Leveraging CNN and Transfer Learning for Vision-based Human Activity Recognition," *2019 29th International Telecommunication Networks and Applications Conference (ITNAC)*, 2019, pp. 1-4, DOI: 10.1109/ITNAC46935.2019.9078016.

