WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Human Activity Recognition Using CNN Techniques

Omkar Radhika[1], Shri N. Naveen Kumar[2]

[1] *(Post Graduate Student, M. Tech (SE) Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad)*
[2] *(Associate Professor, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad)*

A B S T R A C T

As the volume of video and image data grows, the automatic identification of human activity has become a critical task for various applications such as surveillance, health care, smart homes, and interactive systems. Prior to this work, automatic activity recognition in video relied heavily on either wearable sensors or manually engineered features which limited the flexibility and applicability of the approaches in an uncontrolled environment. In this work, we propose a vision-based system for activity recognition based entirely on Convolutional Neural Networks (CNNs) that does not rely on wearable devices and manual feature engineering. Using transfer learning with a pre-trained VGG16 model, we label human activities in video using a CNN-based approach. We demonstrate that our CNN model can effectively classify human activity in real-time, has excellent accuracy and scalability, and is viable for practical applications in uncontrolled real-world environments.

Keywords: Human Activity Recognition, Convolutional Neural Network, Deep Learning, Transfer Learning, VGG16, Video Classification, Computer Vision, Feature Extraction, Action Recognition, Machine Learning, Real-Time Detection, Visual Surveillance, Image Processing, Smart Systems.

1. Introduction

Machine recognition of human activities is a particular area of importance in numerous fields, including surveillance applications, health-care, elder care, and interactive technologies. Human Activity Recognition (HAR) is concerned with the identification of specific physical actions of humans utilizing different input modalities, either sensor or vision related.

Most HAR systems featured wearable sensors that would capture motion (e.g. the Phone or Fitbit) but these systems have many drawbacks such as discomfort to the user, un-generatability beyond the experience design, and lastly dependency on physical devices that are not always as intuitive.

The rise of computer vision and deep learning methods have made where humans can be identified through video, as well as deep learning approaches for vision-based activity recognition a more practical method for behavior recognition instead of existing systems.

Convolutional Neural Networks, a kind of deep learning model, have demonstrated a high potential for visually interpreting static and dynamic input. In this project we will build a CNN to recognize human activities using the video frames instead of traditional sensor inputs. To take advantage of pre-trained knowledge, we will use VGG16 and apply transfer learning to our activity dataset reducing the need of large amounts of data and training time.

2. Literature Review

This section presents an overview of significant research contributions in the field of Human Activity Recognition (HAR), with a focus on vision-based methods that leverage deep learning, particularly Convolutional Neural Networks (CNNs). The shift from sensor-dependent systems to non-intrusive, camera-based recognition systems has been a key trend, driven by the increasing availability of video data and the success of deep learning in computer vision tasks.

[1] Bhandari et al. proposed a novel approach using wrist-worn sensors to automatically detect smoking activities through motion pattern recognition. Their system employed machine learning classifiers to distinguish smoking-related gestures from other hand movements. While their approach achieved high accuracy in controlled environments, its dependency on wearable devices significantly limits its scalability and applicability in natural settings where sensor placement and user compliance cannot be guaranteed. This highlights the need for unobtrusive alternatives such as vision-based recognition.

[2] Yao et al. introduced a device-free HAR system that uses passive RFID signals and compressive sensing techniques to detect human movements. This method eliminates the need for body-worn devices, relying instead on signal distortions caused by user presence and activity. The technique offers a novel direction in HAR by exploring ambient sensing. However, the system's sensitivity to environmental interference and fixed spatial configurations restricts its deployment across varied real-world settings, where dynamic and unstructured environments are the norm.

[3] Lillo et al. addressed the challenge of representing complex activities by proposing a sparse composition framework based on body pose sequences extracted from RGB-D video data. Their model maps activities into a structured sequence of atomic actions, improving recognition accuracy for fine-grained movements. This hierarchical modelling effectively captures both spatial and temporal dependencies. However, the reliance on depth sensors introduces limitations, particularly in outdoor or low-light scenarios where RGB-D accuracy is compromised.

[4] Wharton et al. applied transfer learning to recognize behavioural symptoms in dementia patients using video surveillance data. Their system utilized pre-trained CNN architectures such as VGG16 and Inception to classify patterns of aggression or depression based on postural and facial cues. The model achieved high accuracy in medical monitoring environments and demonstrated the practical use of deep learning in healthcare applications. Nonetheless, it required well-annotated and ethically collected datasets, highlighting the data privacy and labelling challenges in real-world deployments.

[5] Tran et al. proposed Deep 3D Convolutional Neural Networks (C3D) that extend traditional CNNs by learning both spatial and temporal features directly from video volumes. This method marked a significant improvement in modelling the dynamic nature of human activities, as it could capture motion cues over time without manual sequence engineering. Despite the performance gains, the computational demands of 3D CNNs pose a challenge for real-time recognition tasks, making them more suitable for offline analysis or systems with GPU support.

From the above work, it is easy to note that CNN-based vision systems are powerful, effective representations of human behaviours as evidence in visual data. Transfer learning, multi-stream modelling, and spatiotemporal feature extraction are important techniques and systems that have improved recognition methods. Nonetheless, most available approaches either require high computational resources or have been conducted in controlled environments. There remains a need for lightweight and scalable HAR systems, which respect standard video data structures without requiring additional sensor data or pre-processing overhead**.**

3. Defining the problem and objectives

3.1 Problem Definition

Most of the previous work in human activity recognition has been based on wearable sensors or hand-crafted features. Wearable sensors are generally inconvenient and most systems have not generalized well beyond their original subjects and environments. Likewise, hand-crafted features are laborious to engineer and do not intuitively generalize to dynamic and changeable conditions.

Objectives

- Implement a vision-based human activity recognition system using Convolutional Neural Networks.
- Eliminate wearable sensors and use the only video inputs.
- Perform transfer learning from a pre-trained CNN (VGG16) to make using the resource-intensive models efficient and for feature extraction.
- Classify common human activities with as high accuracy as possible.
- Report the model performance using standard performance metrics (e.g. accuracy, confusion matrix, precision-recall).

5. Proposed System

The proposed Human Activity Recognition (HAR) system uses a goal-based, iterative learning process by using Convolutional Neural Networks (CNNs) to apply transfer learning. Instead of developing a deep learning model from scratch, our system is built with a pre-trained VGG16 architecture, significantly reducing training time and computing power needed to train the deep learning model. This approach also enables our model to achieve high accuracy for recognizing human activity with limited amount of labeled data, as additional labelled data are labeled human activity. The model is fine-tuned with the network adapted to identify specific human activities. This system has modular components that begin with the Video Input Module, which can accept pre-recorded videos as well as webcam streaming in real time. The Frame Extraction component harvests several frames from the video stream at specified frames per second, discarding repetitive frames, allowing for important motion characteristics to be preserved. The resulting frames from Frame Extraction go to the Preprocessing component, which maps all frames to 224x224 pixels and eventually normalizes them as VGG16 requires them to the input specifications.

Once the frames are preprocessed, they flow through the CNN Model, where VGG16 consumes the frames and performs feature extraction. The last classification layer produces the predicted human activity label based on the features, which can include activities as walking, sitting, or running. The Prediction Output Module produces the label and confidence score, and provides user interpretation.

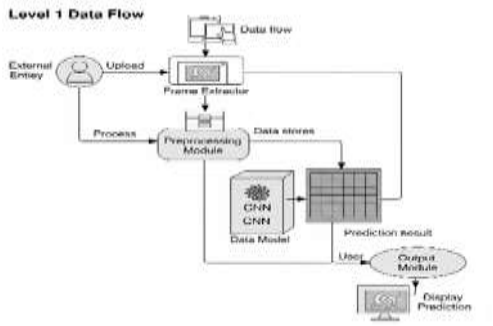

Figure No. 1: Architecture

5. Model and Implementation

The heart of the proposed Human Activity Recognition (HAR) system is a deep learning framework based on Convolutional Neural Networks (CNNs), specifically the VGG16 architecture. We chose this model based on its simplicity, computational efficiency, and very high performance in visual classification tasks. The system is designed to accept video input as input, process the input as frames, and ultimately classify each frame to activities defined in the activity categories of walking, seated, jumping, or running.

CNN and VGG16 Architecture

Convolutional Neural Networks (CNNs) are deep learning models that perform well in image classification and pattern recognition tasks. The use of convolution to extract low-level and high-level features from input images allows for CNNs to be pre-built using existing models and do not require much manual feature engineering. Instead, CNNs learn the most relevant features based on the data, which can include edges, shapes, and motion features.

In this system, we use VGG16 which is a 16-layer deep CNN developed by the Visual Geometry Group at the University of Oxford. The model has 13 convolutional layers and 3 fully connected layers, uses small (3×3) convolutional filters which stack for further complexity to the patterns it can learn, and replaces with our custom classifier. The VGG16 classification output is replaced with a custom classifier taking into consideration updating for human activity recognition. The new classifier includes a Global Average Pooling layer that reduces overfitting along with fully connected layers using the ReLU activation function and Dropout for regularization and a SoftMax layer to give class probabilities.

Also, because we leverage transfer learning, we can preserve the significant ability to extract features of VGG16 all while 'fine-tuning' for our dataset, further reducing training time rapidly with enhanced performance and this is especially useful if data is scarce.

Figure No. 2: VGG16 Architecture using CNN

Input and Preprocessing

First, the system has a video input module, accepting both recorded video and live web cam input. The video must be translated to input for the CNN, so the video was sampled at regular intervals to retrieve frames through the OpenCV package, ensuring temporal consistency and less redundant data while capturing movement and velocity.

Every frame also undergoes preprocessing, rescaling to 224×224 which is the input size for the VGG16 model. Pixel points in each frame were set to the range of [0, 1]. This method allows uniform input distribution. Lighting issues and color space conversions into BGR or BGR to RGB as appropriate maintain consistency with the VGG16 dataset used to train the model, and as required during model development.

Preprocessing is essential for development and providing consistency for increased model stability and improved accuracy. Adequately preprocessing our input data allows the CNN model to learn activity-related characteristics versus non-significant variables, such as size, lighting, or resolution.

Figure No. 3: Input video rescaling

Training and Prediction Output

While training the custom classification layers on a labeled dataset of human activity images, the base VGG16 layers were frozen. This allows the model to take advantage of the highly-trained visual features learned from the pre-training process. Training was completed with the Adam optimizer, using categorical cross-entropy as the loss function. The training ran between 20–25 epochs with a batch size of 32, and a learning rate of 0.0001. To help with generalization, data augmentation was used with horizontal flipping, rotation, zooming, and shifting of images.

After training completed, the system made predictions with its trained model frame-by-frame, either in batch or real-time mode. For each input frame, the model assigned a predicted activity label (e.g., "walking"), as well a confidence score indicating the probability of that class. Predictions were able to be shown in real-time, or logged to be analyzed later, depending on the application.

The model is lightweight and modular. It can be deployed to different platforms such as personal computers, smart surveillance systems, or edge devices. It was capable of inference in real-time while maintaining high accuracy, making it readily applicable to the real-world, including security monitoring, caring for the elderly, and human-computer.

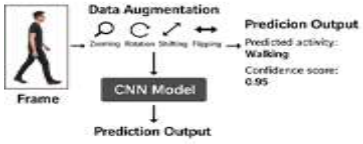

Figure No. 4: Training the model

6. Results

- Running the main Program on local host.

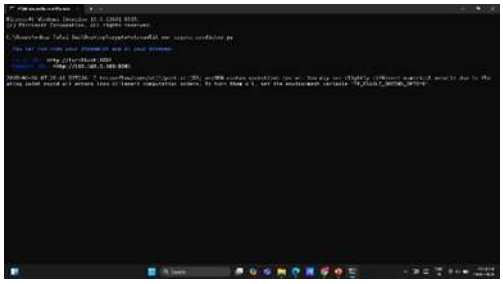

Figure No. 5: Running the Main program

- Execution of the program

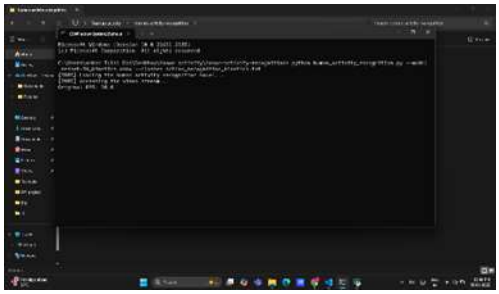

Figure No. 6: Execution of the program

- Output when video is uploaded.

Figure No. 7: Output for input video

- Output of live streaming using laptop camera.

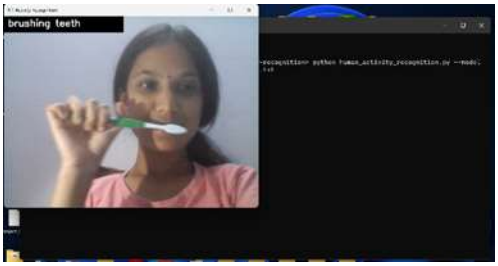

Figure No: 8 Output of the live stream video

7. Conclusion

The research work proposes an efficient and scalable human activity recognition (HAR) system using convolutional neural networks (CNNs), based on the VGG16 model with transfer learning techniques. The HAR system takes video inputs of human activities (e.g., walking, sitting, running, jumping), extracts the frames from the video activities, pre-processes the frames, and passes them to a CNN classifier that has been modified to accommodate smaller frames. Implementing VGG16 as a baseline significantly simplifies and shortens time to train CNNs, while retaining high accuracy even with a small dataset. In addition, the model architecture was developed to be modular, small-footprint, and suitable for real-time application. It is scalable and can be implemented on various platforms, such as personal computers, smart surveillance systems, or edge devices. At the output of the system, the HAR sends reliable labels for each video frame, while delivering predictions along with confidence scores based on predicted probabilities. As such, the HAR is suitable for application in healthcare monitoring, human-computer interaction, and smart environments, among others. The model can demonstrate the capabilities of deep learning in conducting activity recognition from visual data. Some directions for expected future work may include improving temporal dependencies with LSTM networks, adding more classes of activities, and refining the model for mobile and embedded systems.

8.References

[1] B. Bhandari, J. Lu, X. Zheng, S. Rajasegarar, and C. Karmakar, “Noninvasive Sensor-Based Automated Smoking Activity Detection,” IEEE Sensors Journal, vol. 17, no. 15, pp. 4884–4890, 2017.

[2] L. Yao, Q. Z. Sheng, X. Li, T. Gu, M. Tan, X. Wang, and S. Wang, “Compressive Representation for Device-Free Activity Recognition Using RFID Signals,” IEEE Transactions on Mobile Computing, vol. 17, no. 2, pp. 293–306, 2018.

[3] I. Lillo, J. C. Niebles, and A. Soto, “Sparse Composition of Body Poses and Atomic Actions for HAR in RGB-D Videos,” in Proc. European Conference on Computer Vision (ECCV), 2016, pp. 593–607.

[4] Z. Wharton, E. Thomas, B. Debnath, and A. Behera, “A Vision-Based Transfer Learning Approach for Recognizing Behavioral Symptoms in People with Dementia,” Sensors, vol. 19, no. 15, p. 3365, 2019.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks,” in Proc. IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489–4497.

[6] K. Simonyan and A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos,” Advances in Neural Information Processing Systems (NeurIPS), vol. 27, 2014, pp. 568–576.

[7] F. Chollet, “Deep Learning with Python,” Manning Publications, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[9] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” MIT Press, 2016.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in Proc. NeurIPS, 2012, pp. 1097–1105.

[11] OpenCV, “Open Source Computer Vision Library.” Available: https://opencv.org/

[12] TensorFlow, “An End-to-End Open Source Machine Learning Platform.” Available: https://www.tensorflow.org/