Human Activity Recognition with Open CV and Deep Learning

Y.Prabhu Prakash, P.Dileep, P.Darwin
Electronics and Communication Engineering, Godavari Institute of Engineering & Technology (A).

Abstract: **Human Activity Recognition (HAR) plays a crucial role in various applications such as surveillance, healthcare, and human-computer interaction. In this paper, we propose a new HAR system via Convolution Neural Networks (CNN) which is one of deep learning algorithms. The proposed system encompasses a comprehensive dataset comprising 400 diverse human activities, enabling a robust and versatile model. Leveraging the power of OpenCV (Open Source Computer Vision Library) and Deep Learning techniques, the system aims to automatically identify and classify various human actions in real-time. The dataset encompasses a wide spectrum of activities, ensuring the model's ability to generalize across diverse scenarios. The 3D CNN model used in this model is RESNET-34. The activities include both routine daily tasks and anomalous behaviors, providing the system with the capacity to identify unusual or suspicious actions. The extensive dataset also contributes to the adaptability of the model in different environments and cultural contexts. The process involves capturing video data, preprocessing it using OpenCV to extract relevant features, and feeding these features into a Deep Learning model. The model, trained on a dataset of diverse human activities, learns to associate specific patterns with different actions. This enables the system to predict and classify ongoing activities, ranging from simple gestures to complex movements. We have achieved the average recognition accuracy of 89.99% for the activities.**

Keywords: **Convolutional neural networks (CNN), Human activities recognition(HAR), OpenCV, Resnet.**

I. INTRODUCTION

Human Activity Recognition (HAR) is a field of study that involves the use of computer vision and deep learning techniques to identify and classify human activities based on data collected from cameras or video feeds. Human Activity Recognition involves analyzing and interpreting data from video feeds to recognize different activities performed by individuals. These activities could include actions like walking, running, sitting, or even more complex activities like playing sports.

For video-based recognition, a dataset containing labeled video clips of different activities is required. Each video clip should be labeled with the corresponding activity it represents. This involves extracting frames from the videos and possibly resizing them to a standard size. Additionally, the data may need normalization to ensure that all features have a consistent scale. Convolutional Neural Networks (CNNs) are commonly used for image-based recognition tasks. The primary motivation behind this choice is the limitations of certain sensors, like GPS receivers, in outdoor environments. While many sensors, such as body inertia sensors, can record human actions, their effectiveness may be limited outdoors.

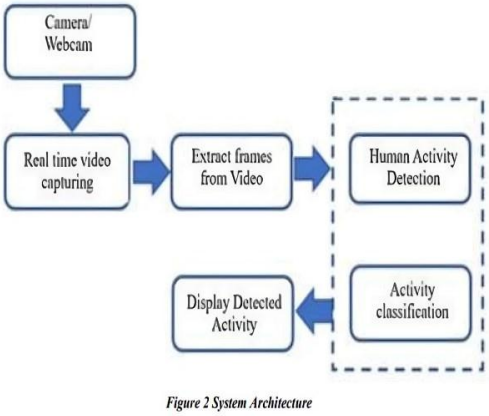

Fig.1 System Architecture

ResNet-34 is an artificial neural network (ANN) algorithm inspired by the structure of pyramidal cells in the cerebral cortex.[1]It utilizes skip connections, allowing it to "jump over" some layers in the neural network. The algorithm is specifically designed with two and three-layer skips, incorporating non-linearities (ReLU) and batch normalization. These skip connections, also referred to as HighwayNets, can be recalled using an additional weight matrix. The procedure of skipping layers is similar to DenseNets when multiple levels of parallel skips are involved. Dense Nets use skipping of multiple neural network layers.[2]Skipping levels in residual networks,

whether HighWayNets or Dense Nets, helps resolve the issue of vanishing gradient. This is done by reusing previous layer activations until the adjacent layers are capable of learning their weights. Non-residual networks, which are not ResNet, behave like plain networks. The main purpose of skipping levels in residual networks is to address the vanishing gradient problem by reusing previous layer activations until the adjacent layers can learn their weights. This is crucial for effective learning in deep neural networks.[3] The paper emphasizes that the best use of ResNet occurs when a single layer is overskipped. In simpler terms, this means using all linear layers as intermediate layers.[4] If this is not feasible, a specific weight matrix should be created for all skipped connections, avoiding the use of Dense Net and opting for High way Net instead. Skipping multiple layers in ResNet is effective for simplifying the network during initial training stages. This speeds up the learning phase by reducing the impact of gradients due to fewer layers in the propagation process.[5] The skipped layers are gradually reconstructed through the feature space learning process, aiding in the accurate recognition of human activities.[6]

II. METHEDOLOGY

The methodology for human activity recognition with OpenCV and deep learning typically involves several key steps. It Implementation involves two primary processes training and recognition.[7] To initiate the training process, we select a temporal position in a film to generate training samples using sampling. If the chosen video clip is shorter than sixteen frames, we loop around it until obtaining a sixteen-frame segment. Subsequently, we determine a spatial position and scale according to the requirements, resizing the samples to 112 X 112 pixels. During the training of the Resnet-34 model from scratch, the initial learning rate is set to 0.1 and later reduced by a factor of 0.1 after the validation loss saturates. Moving on to the recognition phase, a loop begins over the frames, initializing a batch of frames to be passed to the neural network. We populate the batch from the video stream, resizing them to a width of 400 pixels while maintaining aspect ratios.[8]This approach facilitates building a batch of multiple images for the human activity recognition network, allowing it to leverage spatiotemporal information. The dataset used for training is the Kinetics human action video dataset, encompassing 400 classes of human activities. Each action has 400 or more films, each lasting around a tenth of a second, extracted from various shutterstock video clips. It is important to note that the success of the methodology relies on the choice of deep learning model, proper training, and thoughtful preprocessing of video data, demonstrating the integration of both OpenCV for video manipulation and deep learning for activity recognition. Feature extraction techniques are then utilized to represent relevant information, such as body joint positions or motion descriptors. Machine learning models, including classifiers or deep neural networks, are trained on labeled datasets to recognize specific activities based on the extracted features. Post-processing steps, such as temporal smoothing or filtering, may be applied to improve the accuracy and robustness of the recognition system. Finally, the recognized human activities are visualized or logged for further analysis or real-time applications, demonstrating the effectiveness of the OpenCV-based methodology in understanding and interpreting human actions from visual data. The dataset used for training, such as the Kinetics human action video dataset, contributes to the model's ability to generalize across a diverse range of human activities, encompassing various classes and interactions such as cleaning floor, frying vegetables, weaving basket, chooping wood, extinguishing fire as well as human-human interactions such as hand shaking, hugging etc.[9]

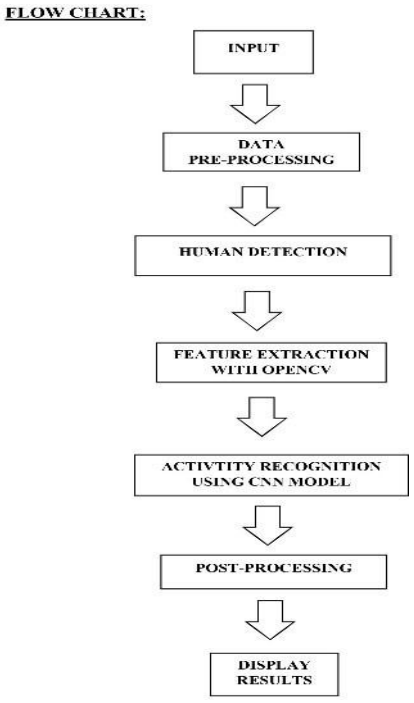

Fig.2 Flow Chart

The specified sample has a duration of 16 seconds and a frame rate of 112 frames per second, resulting in a total of 1792 frames. Random data is generated for five distinct activities— "Walking," "Running," "Sitting," "Standing," and "Jumping." The subsequent line plot visualizes the intensity of each activity over time, with simulated variations based on a normal distribution.

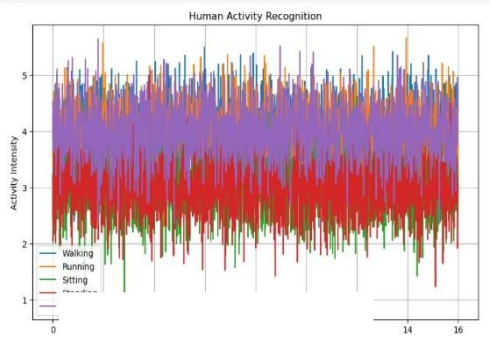

Fig.3 Regular Activities

The accuracy percentage of human activity recognition. It measures the accuracy of recognizing various activities, ranging from 75% to 95%. The activities listed at the bottom include walking, running, sitting, standing, jumping, smoking, playing chess, feeding fish, making jewelry, and skipping rope. Each activity presumably corresponds to a specific point on the graph, although the image does not provide explicit connections between the activities and their respective accuracy percentages. The graph's y-axis is labeled 'Accuracy Percentage (%)', divided into increments of 2.5%, starting from 75.0 and ending at 95.0. The x-axis, while not explicitly labeled, seems to represent the different activities being recognized. Text overlay on the graph provides additional context, with percentages ranging from 75% to 95% marked out on the y-axis. The x-axis is populated with the names of the activities being recognized, which include a mix of physical activities like walking and running, as well as more sedentary ones such as playing chess or making jewelry.

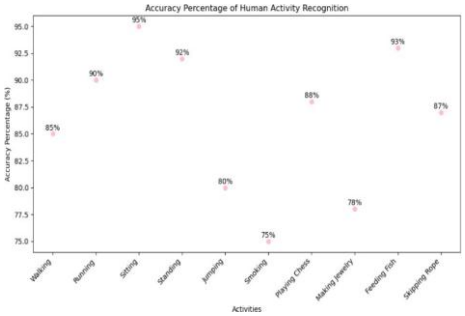

Fig.4 Accuracy (%) of HAR

This image is a graphical representation that appears to illustrate the concept of "Human Activity Recognition". It's a quantitative chart that showcases different types of human activities and their corresponding number of frames. The activities listed include making jewelry, skipping rope, feeding fish, smoking, playing chess, jumping, standing, running, sitting, and walking. Each activity is represented on the Y-axis while the X-axis seems to display the number of frames ranging from 0 to 1000. The dominant colors in the image are white, with a subtle accent of a blue hue, possibly representing the lines or bars of the graph. Despite the numerical data and the precision of the graph, no human faces, celebrities, or landmarks are identified in the image. It seems to be a screenshot of a display, perhaps from an analytical software or digital tool, with text and lines in a clear, readable font. The image does not contain any explicit adult, racy, or gory content, making it suitable for a wide array of audiences. Overall, this image provides a visual representation of how different human activities can be quantified and analyzed through the number of frames.

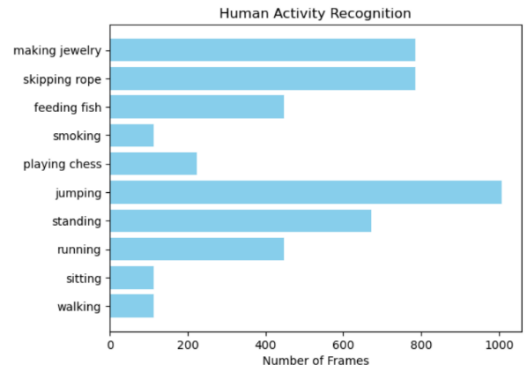

Fig.5 Numbers of frames are Occurring

III. RESULT AND DISCUSSION

The trained model gave an accuracy of 89% on the kinetic dataset. We observed that the accuracy was very high for activities like walking, running,sitting, standing, jumping etc. but it was reduced considerably for activities like cooking, doing yoga, etc., since there are several ways of performing these activities. For further improvement of results, we can more detailed dataset which separates the different yoga asanas into different labels. We observe that datasets with more detailed class labels give better results. So, instead ofusing the broad term cooking, splitting the class into different labels like cooking rice, boiling water,playing chess, feeding fish, making jewellery,hula hooping,smoking etc. will certainly lead to better results.

Fig.6 Playing Chess

Fig.7 Feeding Fish

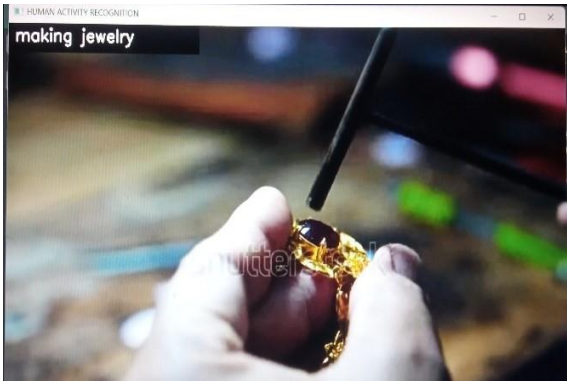

Fig.8 Making jewellery

Fig.9 Skipping rope

Fig.10 Smoking

The dataset contains a single activity in each entry, examples like 2 people in the same frame performing different activities were not considered. For such entries, first performing some video processing to determine person of interest in the frame and then using the model to determine different activities will be sufficient.

IV. CONCLUSION

In this paper Human Activity Recognition System, we propose a model trained using Convolutional Neural Network (CNN) with spatiotemporal three-dimensional kernels on the Kinetic dataset. This model demonstrates effective recognition of nearly 400 human activities with a commendable level of accuracy. The developed system serves various practical applications, such as automatically categorizing a video dataset, facilitating training and supervision of new employees in task performance, verifying food service workers, and monitoring patrons in bars/restaurants to ensure quality service. As part of future work, expanding the dataset to cover more than 400 activities is considered to enhance the system's versatility. Additionally, our observations indicate that increasing the number of samples for each activity in the dataset significantly improves the system's performance

ACKNOWLEDGMENT

We the team members of the research project sincerely want to thank our guide Assistant Professor Mr.G.V.Vinod and the reputed Electronics and Communication Department of Godavari Institute of Engineering and Technology, Rajamahendravaram, India. For their helpful and motivating encouragement and the must needed support for the completion of this project work by

providing the golden opportunity in the form of Major Project.

REFERENCES

[1] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian (2015-12- 10). "Deep Residual Learning for Image Recognition".

[2] Srivastava, Rupesh Kumar; Greff, Klaus; Schmidhuber, Jürgen (2015-05-02). "Highway Networks".

[3] Huang, Gao; Liu, Zhuang; Weinberger, Kilian Q.; van der Maaten, Laurens (2016-08-24). "Densely Connected Convolutional Networks

[4] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici,B. Varadarajan, and S. Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. arXiv preprint, arXiv:1609.08675, 2016.

[5] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the International Conference on Computer Vision ICCV), 2017.

[6] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory- pooled deep-convolutional descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 43054314, 2015.

[7] BishoySefen, Sebastian Baumbach et.(Human Activity Recognition Using Sensor Data of Smart phones and Smart watches ICAART 2016).

[8] un X, Chen C, Manjunath BS , ―Probabilistic motion para-meter models for human activity recognition,‖ In: Proceedings of 16th international conference on pattern recognition, pp 443–450.

[9] Kwon W, Lee TW, ― Phoneme recognition using ICA- based feature extraction and transformation, Signal Process 84(6):1005–1019, 2004.

[10] Lee SI, Batzoglou S, ―Application of independent compo-nent analysis to microarrays,‖ Genome Biol 4(11):R76.1–21, 2003.