# RECOGNITION OF HUMAN ACTIONS BASED ON DEEP CONVOLUTIONAL NEURAL NETWORK USING POSTURES AND DEPTH MAPS

**[1] Pratheeksha R, [2] Rakshitha G K, [3] Vijay Adithya B K, [4] Anusha Preetham**

*[1][2][3] UG Students*

*[4] Asst. Professor, Department of Information Science and Engineering*

*Dept. of Information Science & Engineering*

*DSATM, Bangalore-560082*

## ABSTRACT

*The Human Action Recognition is one of the moat has become a significant space in the PC vision and furthermore has become an essential need for different PC applications that utilize individuals conduct.We can see the use of HAR in various fields of our everyday lives among which many come under surveillance (For traffic surveillance, hospital or bank activities,etc).As we have already seen or come through different methodologies already in existence for recognizing different human activities depending on postures.But there are very few that consider live input for action recognition.These are frameworks which are competent to perceiving the unpredictable human activity designs with a contribution from advanced camera and sensors. This implementation paper gives a useful and precise computation for various and distinct activity recognition. This paper presents a completely amazing part extraction and exhaustive gathering methodologies of the human exercises and getting ready measure. It essentially focuses to arrange the human movement plans with picture recuperation from the video input.The main difference between our system and already prevailing system is that out system is implemented on Deep - Convolutional neural network(D-CNN) by achieving an accuracy of 98.97% through image input,97.43 through video input and 79.07% through live input.*

## I INTRODUCTION

Inside an enormous scope of uses in PC vision, Human Action Recognition has gotten perhaps the most appealing re-search fields. Ambiguities in perceiving activities doesn't just come from the trouble to characterize the movement of body parts, yet additionally from numerous different difficulties identified with genuine issues, for example, camera movement, dynamic foundation, and terrible climate conditions. There has been little exploration work in reality states of human activity acknowledgment frameworks, which urges us to truly look in this application space. Albeit a plenty of hearty methodologies have been presented in the writing, they are as yet lacking to completely cover the difficulties. To quantitatively and subjectively think about the presentation of these strategies, public datasets that present different activities under a few conditions and requirements are recorded. The proposed framework targets ordering different human activities utilizing three info techniques [live video input, pre-stacked video input, picture input] utilizing Convolutional Neural Networks calculation. As a continuously motorized approach, the CNN strategy utilizes picture data as information data and can be directly named yield. The machine will recognize the picture of various human exercises in qualities with different targets and measurements while noticing the interruption in the yield because of high goal of pictures and vacillation of sizes and shapes, in this manner doing the helpful side of the arrangement work and improving the exactness of grouping steps. Numerous strategies for distinguishing human activities exist, yet we will zero in on Convolutional neural organizations that uses profound learning methods.

AI is a part of computerized reasoning (AI) zeroed in on building applications that gain from information and improve their precision over the long haul without being modified to do as such. In information science, a calculation is a grouping of factual preparing steps. In AI, calculations are 'prepared' to discover pat-terns and highlights in enormous measures of information to settle on choices and forecasts dependent on new information.

The better the calculation, the more precise the choices and expectations will become as it measures more information. Today, instances of AI are surrounding us. Advanced colleagues search the web and play music in light of our voice orders. Sites suggest items and films and melodies dependent on what we purchased, watched, or paid attention to previously. Robots vacuum our floors while we improve our time. Spam indicators prevent undesirable messages from coming to our inbox-es. Clinical picture examination frameworks help specialists spot tumors they may have missed. What's more, the primary self-driving vehicles are taking off. We can hope for something else. As large information continues getting greater, as figuring turns out to be all the more remarkable and moderate, and as information researchers continue to foster more competent calculations, mama chine learning will drive more noteworthy and more prominent effectiveness in our own and work lives.

Profound learning is a subset of AI (all profound learning is AI, yet not all AI is profound learning). Profound learning calculations characterize a fake neural organization that is intended to become familiar with the manner in which the human cerebrum learns. Profound learning models require a lot of information that pass through different layers of computations, applying loads and inclinations in each progressive layer to constantly change and improve the out-comes. Profound learning models are normally unaided or semi-managed. Support learning models can likewise be profound learning models. Particular sorts of profound learning models—including convolutional neural organizations (CNNs) and repetitive neural organizations (RNNs)— are driving advancement in regions like PC vision, common language handling (counting discourse acknowledgment), and self-driving vehicles. A CNN is a kind of DNN comprises of various secret layers, for example, convolutional layer, RELU layer, Pooling layer and completely associated a standardized layer. CNN shares loads in the convolutional layer lessening the memory impression and expands the presentation of the organization. The significant highlights of CNN lie with the 3D volumes of neurons, nearby availability and shared loads. An element map is created by convolution layer through convolution of various sub areas of the info picture with a learned piece. Then, at that point, anon-straight initiation work is applied through ReLu layer to improve the intermingling properties when the mistake is low. In pooling layer, an area of the picture/include map is picked and the pixel with greatest worth among them or affirm age esteems is picked as the delegate pixel so a 2×2 or 3×3 lattice will be decreased to a solitary scalar worth.

This outcomes in a huge decrease in the example size. Once in a while, the conventional Fully-Connected (FC) layer will be utilized related to the convolutional layers towards the yield stage. In CNN engineering, generally convolution layer and pool layer are utilized in some blend. The pooling layer typically does two sorts of tasks viz. max pooling and implies pooling. In mean pooling, the normal area is determined inside the component focuses and in max pooling it is determined inside a maxi-mum of highlight focuses. Mean pooling decreases the mistake brought about by the local size restriction and holds foundation in-development. Max pooling lessens the convolution layer boundary assessed mistake brought about by the mean deviation and henceforth holds more surface data.

During our study research, we experienced the accompanying systems been utilized. Right off the bat, Depth Motion Image descriptor (DMI) enables the significance aides of a development to clutch the changing through and through of human development. The following technique being MJD (Moving Joints Descriptor) which shows human joints movement over the long haul by utilizing the 3D round facilitates.

## II EXISTING METHODOLOGIES

From the survey carried out, we have come to observe that,

In [5], a method is proposed which is compact and effective nonetheless easy methodology to code spatiotemporal info carried in 3D skeleton sequences into multiple 2nd pictures, brought up as Joint mechanical phenomenon Maps (JTM), and ConvNets are adopted to use the discriminative options for real-time action recognition. The proposed methodology has been evaluated on 3 public benchmarks, i.e. Different public datasets avail-able online and achieved the state-of-the-art results. [6] In this paper, another technique for 3D activity acknowledgment with

skeleton successions (i.e., 3D directions of human skeleton joints). The proposed technique at first changes each skeleton course of action into three fastens each comprising of a couple of edges for spatial common segment getting the hang of using significant neural associations. Each clasp is produced from one channel of the barrel shaped directions of the skeleton succession. Each casing of the created cuts speaks to the transient data of the whole skeleton succession, and consolidates one specific spatial connection between the joints. The whole clasps incorporate various edges with various spatial connections, which give valuable spatial primary data of the human skeleton. We propose to utilize profound convolutional neural organizations to learn long haul transient data of the skeleton arrangement from the casings of the produced clasps, and afterward utilize a Multi-Task Learning Network (MTLN) to together handle all edges of the created cuts in corresponding to consolidate spatial underlying data for activity acknowledgment. Exploratory outcomes plainly show the viability of the proposed new portrayal and highlight learning technique for 3D activity.

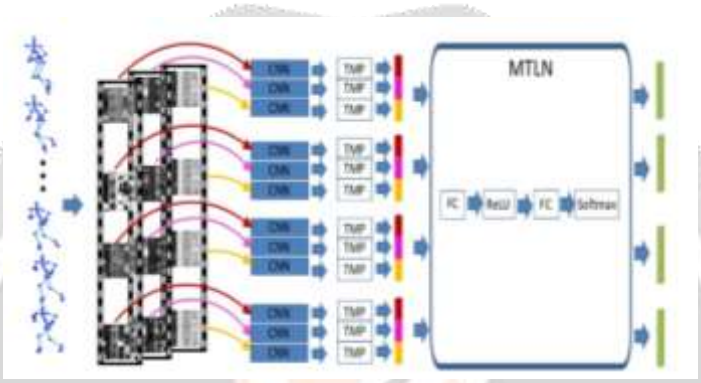


Fig. 3. Existing MTLN model.

[4] This method proposes to associate end-to-end gradable architecture for skeleton primarily based action recognition with CNN. At first, we may show the skeleton sequence as a matrix by joining the joint coordinates in each instance and organizing these vector pictures during the written account order. Then the extracted matrix is simplifying into an image and is made normal to make use of the length of the variable to be a setback. The final picture is uploaded into a CNN model for feature extraction and recognition. For specific structure of each image, the simple max-pooling plays a very vital role on abstraction feature selection, along with temporal frequency adjustment, this may get a lot of differentiated joint data for various actions along with, addressing the variable-frequency problem. [8] This paper puts forward, approaches for consolidating various wellsprings of information in profound learning. In the first place, we propose high-light increase, where we use a collaborator, hand-made, feature (for instance optical stream) to perform spatially moving smooth gating of center CNN's may have maps included. Second, we present a spatially changing multiplicative combination strategy for joining multiple CNNs prepared on various sources that outcomes in vigorous forecast by enhancing or smothering the component activations dependent on their under-standing. We test these procedures in the setting of movement affirmation where information from spatial and transient signs is useful, getting results that are essentially indistinguishable with state of the art methods and defeat systems using just CNNs and optical stream features. [9] This paper depicts an uninhibitedly accessible dataset, named UTD-MHAD, which comprises four transiently synchronized information modalities. These modalities incorporate RGB recordings, profundity recordings, skeleton positions, and inertial signs from a Kinect camera and a wearable inertial sensor for a far-reaching set of 27 human activities. Test results are given to show how this information base can be utilized to contemplate combination moves toward that include utilizing both profundity camera information and inertial sensor information. This public domain dataset is good for multimodality research exercises being directed for human activity acknowledgment by different examination gatherings. [10] proposed a temporal division and classification technique that represents progress designs between occasions of interest. We apply this technique to naturally recognize striking human activity occasions from recordings. A discriminative classifier (e.g., Support Vector Machine) is utilized to perceive human activity

occasions and a proficient unique programming calculation is utilized to mutually decide the beginning and finishing transient fragments of perceived human activities. The critical contrast from past work is that we present the displaying of two sorts of occasion progress data, specifically occasion change sections, which catch the event designs between two back to back occasions of inter-est, and occasion progress probabilities, which model the progress likelihood between the two occasions. Test results show that our methodology altogether improves the division and acknowledgment execution for the two datasets we tried, in which particular progress designs between occasions exist. [18] This paper presents of profundity sensors, for example, Microsoft Kinect have driven exploration in human activity acknowledgment. Human skeletal information gathered from profundity sensors pass on a huge measure of data for activity acknowledgment. While there has been impressive advancement in real life acknowledgment, generally existing skeleton-based methodologies disregard the way that not all human body parts move during numerous activities, and they neglect to consider the ordinal places of body joints. Here, and spurred by the way that an activity's class is controlled by neighborhood joint developments, we propose a cuboid model for skeleton-based activity acknowledgment. In particular, a cuboid organizing system is created to coordinate the pairwise relocations between all body joints to get a cuboid activity portrayal. Such a portrayal is all around organized and permits profound CNN models to center investigations on activities.

**THE GAPS ENCOUNTERED DURING THE SURVEY OF HAR USING CNN TECHNOLOGY**

The gaps encountered during the survey of human action recognition using convolutional neural networks technology are as follows:

1. Despite the fact that CNN is equipped with great feature extraction and classification in many of the computer vision problems, the CNN model is unable to classify the actions correctly specifically when the input images do not equip with discriminative features.

2. The current models all portray the exercises reliant upon the overall spatial and transient information found in the skeleton progressions. This requires the uproar flow in different bits of a comparable gathering to be dependable. Therefore provoking the affirmation rate been cleaved down, if the data bumble of neighborhood segments in the information groupings is included.

3. It is understood that all of the association meets to some substitute neighborhood minima, whether or not every association were to be set up on a comparative data technique. The show increases when joining different associations alongside clear late mix approach in light of each close by minima having an insignificantly one of a kind data. Thus, there is a need to multiplicatively unite various CNNs to orchestrate the photos.

4. It is understood that all of the association meets to some substitute neighborhood minima, whether or not every association were to be set up on a comparative data system. The display increases when joining different associations alongside clear late blend approach in view of each close by minima having a hardly remarkable data. Thusly, there is a need to multiplicatively unite various CNNs to organize the photos.

## III OBJECTIVE OF THE PROJECT

1. The main aim of our project is to achieve greater accuracy than the existing methodologies in activity recognition.And have managed to fullfill it to a greater extent.

2. In this we have considered three different input types namely,image input pre recorded video input and live video input.

3. We have trained almost 4000 vivid actions using the data set from KAGGLE(UTF-80).The Accuracy can be further increased by training further more actions.

## IV SYSTEM OVERVIEW

In our model we have basically implemented our model based on the budding technology DEEP-CONVOLUTED NEURAL NETWORK (D-CNN) where the images are trained and recognized using different frames of image.The user feeds the system with input and the input is divided into frames and then the action is recognized using various inbuilt packages and then the result is displayed on the screen.

Then the results are tested for various input sets and accuracy is recorded.

## V SYSTEN IMPLEMENTATION

Each project is divided into a number of processes or phases, each phase having its own identity and characteristics, the respective phases are:


FIG 4:SYSTEM DESIGN

**Image Acquisition (Input Image)**

Picture obtaining in picture preparing can be comprehensively characterized as the activity of recovering a picture from some source, normally it is an equipment based source, so it tends to be gone through whatever cycles need to happen a while later. Performing picture procurement during the time spent picture preparing is consistently the initial phase in the work process succession on the grounds that, without a picture, no handling is conceivable. The picture that is obtained is totally natural and is the aftereffect of scanner which was utilized to create it, which can be vital in certain fields to have a predictable gauge from which to work. One of a definitive objectives of this interaction is to have a wellspring of info that works inside such controlled and estimated rules that a similar picture can, if fundamental, be almost consummately replicated under similar conditions so irregular elements are simpler to find and kill.

**Preprocessing**

Picture pre-handling is done to fortify or escalate a portion of the highlights of picture significant for future examination and preparing. Clamor from the isolated picture is killed utilizing middle channel. Middle channel is essentially subject to a moving window over the whole picture and ascertaining the resultant pixel esteem as the middle worth of the splendor esteem in the current window. The subsequent smoothed picture channels are reestablished. Another pre-preparing embraced was to standardize the size of different cash notes by keeping a similar perspective proportion. The viewpoint proportion can be characterized as the proportion of the width of the note to that of the stature of the net. Dissimilar to the size of the picture the angle proportion of note of specific division is free of the separation from which the photograph of the picture has been taken. In the proposed approach, we kept the angle proportion of (66, 166). Picture is then changed over into Grayscale as Image handling utilizes the idea of „comparing" areas in a picture. Examination in Grayscale includes straightforward scalar mathematical administrators (+, - ). In any case, separate tones are required, the strategies are somewhat more mind boggling. Normally, to get great outcomes, some sort of Vector distinction is required. This is computationally more unpredictable, and still doesn't give ensured better outcomes. Power information is normally adequate. Grayscale (for example power) is typically adequate to recognize edges. As

seen here, handling tone is intricate, and Grayscale gives a path of least resistance. Nonetheless, the facts demonstrate that shading picture handling can give better outcomes. The crucial step is guaranteeing right computations. Picture Binarization is likewise accomplished dependent on the limit. im2bw() work changes over the grayscale picture to a twofold picture.

**Image Segmentation**

It decides district limits in a picture. It can investigate a wide range of ways to deal with a picture Segmentation and thresholding. Ideal Global Thresholding:

1. An edge is supposed to be all around the world ideal if the quantity of misclassified pixels is least.

2. Histogram is bimodal (object and background).

3. Ground truth is known OR the histograms of the item and the foundation are known.

**Feature Extraction**

Highlight extraction a kind of dimensionality decrease that productively addresses fascinating pieces of a picture as a reduced element vector. This methodology is helpful when picture sizes are huge and a diminished component portrayal is needed to rapidly finish undertakings, for example, picture coordinating and recovery. A few highlights of a picture are: Size or Area. Each group varies from one another in the size boundary. Subsequently size can be utilized as an element for cash acknowledgment. In any case, the significant impediment of this component is that the size of the picture differs relying upon the separation from which photograph of the picture has been taken. To conquer this issue another boundary named angle proportion was utilized to group the divisions.

**Comparison**

In our correlation the highlights removed from the pictures of the money notes assumes an extremely essential part. In realities, it is the examination of the highlights that empowers us to separate phony notes from the genuine ones. To analyze the exhibition, we have sectioned the picture and afterward we eliminate from a double picture every associated segment (protests) that have less than P pixels, creating another parallel picture. The above advance is rehashed multiple times to get the twofold picture which can measure up. Then, at that point we look at the two pictures and store the distinction.

## VI METHODOLOGY

The framework of the proposed action recognition method is :we use two types of data for human action representation, depth maps, and body postures. Each depth map frame is associated with the body postures. Each of the two inputs is transformed to a descriptor that assembles the input sequence in one image in order to provide an informative description of the action. Namely, DMI for depth maps and MJD for body joints. The DMI descriptor captures the changing in depth of the action during the body motion. The MJD descriptor which inspired from the nature of the human body joints movement around a fixed point to capture,the joints direction and the changing in the joint position. The MJD descriptor overcome the lack of side views in the DMI descriptor Three CNN models of the same structure are trained and tested with the two descriptors in a way that one model takes two descriptors as input and each of the two other models takes only one descriptor. The reason behind this assumption is to exploit the power of CNN for extracting features from the two descriptors in different ways with multiple channels for the sake of improving the classification accuracy.

We propose several score fusion operations to get a high score of the accuracy prediction by combining the outputs of the three models. The model training and testing are performed on three action datasets that contain both depth images and posture data.

## VII RESULTS

Fig 5 Basic landing pages

Fig 6 Recognizing image inputs

Fig 7 Recognizing video inputs

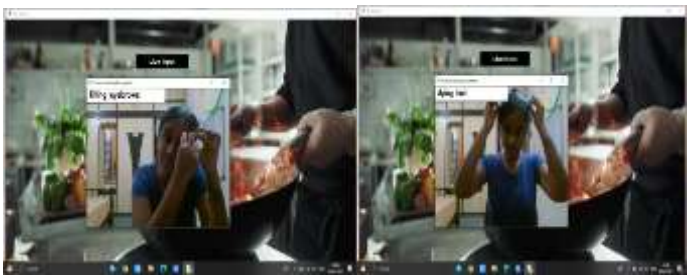

Fig 8 Live input Recognition

## VIIICONCLUSION

Human action acknowledgment is indispensable for various PC vision applications that demand information of people lead, including reconnaissance for public security, human cooperation applications and mechanical innovation. Regardless, action affirmation in concealed pictures is trying assignment on ac-count of a couple of segments, for instance, complex establishment, edification assortment, and clothing tone, which make it difficult to part the human body in every scene. For this we have trained the CNN model with UCF-101. The CNN model is trained with 4000 videos datasets in 80:20 ratios, which have proven to be having an overall maximum accuracy of 98%. The most efficient of all being considered to be the usage of D-CNN's for action recognition, though there are a few drawbacks even in this method since improper inputs are not provided during the input.

## REFERENCES


1 W. Chi, J. Wang, and M. Q.-H. Meng, "A gait recognition method for human following in service robots," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017.

2 J. Yu and J. Sun, "Multiactivity 3-d human pose tracking in incorporated motion model with transition bridges," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017.

3 G. Liang, X. Lan, J. Wang, J. Wang, and N. Zheng, "A limb-based graphical model for human pose estimation," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017.

4 Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on. IEEE, 2015, pp. 579–583.

5 P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," arXiv preprint arXiv:1612.09401, 2016.

6 Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," arXiv preprint arXiv:1703.03492, 2017.

7 ——, "Skeleton optical spectra based action recognition using convolutional neural networks," arXiv preprint arXiv:1703.03492, 2016.

8 E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016, pp. 1–8.

9 C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, 2015, pp. 168–172.

10 Y. Kim, J. Chen, M.-C. Chang, X. Wang, E. M. Provost, and S. Lyu, "Modeling transition patterns between events for temporal human action segmentation and classification," in Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, vol. 1. IEEE, 2015, pp. 1–8.

11 C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on. IEEE, 2015, pp. 1092–1099.

12 J. Koushik, "Understanding convolutional neural networks," arXiv preprint arXiv:1605.09081, 2016.

13 J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, and G. Wang, "Recent advances in convolutional neural networks," arXiv preprint arXiv:1512.07108, 2015.

14 E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016, pp. 1–8.

15 ] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "Rgb-d-based action recognition datasets: A survey," Pattern Recognition, vol. 60, pp. 86–105, 2016.

16 C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, 2015, pp. 168–172.

17 Han Zhao, Xinyu Jin, "Human Action Recognition Based on Improved Fusion Attention CNN and RNN", 2020 5th International Conference on Computational Intelligence and Applications (ICCIA).

18 Kaijun Zhu, Ruxin Wang, Qingsong Zhao, Jun Cheng, and Dapeng Tao, "A Cuboid CNN Model with an Attention Mechanism for Skeleton-based Action Recognition".