Recognition of human actions based on deep convolutional neural networks using postures and depth maps

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

[1] *Student, Department of Information Science &engg, DSATM, Karnataka,India*
[2] *Student, Department of Information Science &engg, DSATM, Karnataka,India*
[3] *Student, Department of Information Science &engg, DSATM, Karnataka,India*
[4] *Asst. Professor, Department of Information Science &engg, DSATM, Karnataka,India*

ABSTRACT

In this survey paper, we have gathered information about various existing technologies in the area of Human Action Recognition Using Convolutional Neural Network with Human Postures and Depth Maps. Amongst all the available technologies we have focused on a method where we use MJD (moving joints descriptor) and DMI (depth motion image) for action recognition. Profundity movement pictures depict the overall action appearance by gathering all profundity guides of the action throughout an opportunity to make a uniform depiction that can portray every action with its own specific appearance from the front view. It gets the changing inside and out of the moving body parts. The DMI depiction give unquestionable features to every movement which encourages the component extraction task for the CNN model. Profundity movement pictures depict the overall action appearance by gathering all profundity guides of the action throughout an opportunity to make a uniform depiction that can describe every action with its own specific appearance from the front view. It gets the changing top to bottom of the moving body parts. The DMI depiction give unquestionable features to every action which encourages the element extraction task for the CNN model.

I INTRODUCTION

The Human Action Recognition has become a significant do-primary in the PC vision and furthermore has become an imperative need for different PC applications that utilize individuals' conduct. For the ramifications of an activity acknowledgment framework there is a necessity of insignificant calculations, including a wide scope of utilizations like the gesture based communication acknowledgment, reconnaissance, and video investigation. These are systems which are capable to recognizing the complex human action patterns with an input from digital camera and sensors. This study paper gives a productive and exact calculation for human activity acknowledgment. This paper portrays a powerful component extraction and thorough grouping of the human activities and preparing measure. It basically centers to order the human activity designs with picture recovery from the video input.

Based upon the survey, we observe that:

- To upgrade the shortcomings of the utilization of one sort of records for movement acknowledgment, two movement portrayals are utilized. Profundity map outline and body joint portrayal. The proposed constitution joints outline is invigorated by utilizing the way that the human build joints pass to cover the joints course notwithstanding the adjusting in joints position.

- An all-around planned CNN model is prepared exceptionally to ex-plot highlights from the two sorts of move portrayal, taking the calculation time in thought by utilizing "Network in Network" structure. Three channels of the model are utilized to separate highlights from different information.

- Mix errands between gauge consequences of the three CNN coordinates are proposed to adorn the assumption precision. The proposed technique presents a versatility in picking the best way to deal with

orchestrate the development by strategies for two sorts of data, three CNN channels, and various mix assignments.

- A goliath amount of training records is one of the critical accomplishment of a CNN life sized model forecast exactness. Because of the absence of a monster RGB-D movement awareness dataset, the utilization of two movement portrayals assists with upgrading the examining methodology on a little amount of information.

During our survey research, we encountered the following methodologies been used. Firstly, Depth Motion Image descriptor (DMI) empowers the profundity guides of a movement to hold onto the modifying top to bottom of human movement. The next procedure being MJD (Moving Joints Descriptor)

which shows human joints motion over time by using the 3D circular coordinates.

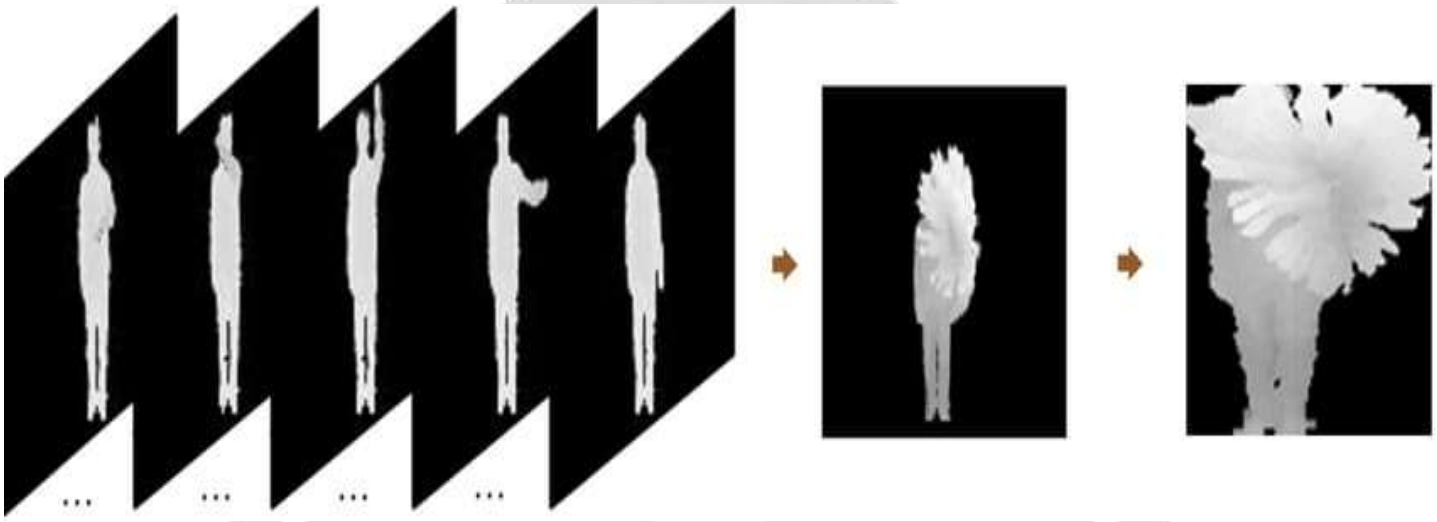

Fig. 1. Depth Motion Image descriptor (DMI).

Depth Motion Image (DMI): -We can represent this technology using the example of a person drawing circle (action being performed) from the given dataset, middle: DMI (Depth Motion Image), left: depth map sequences, right: Cropping Region of Interest(ROI).

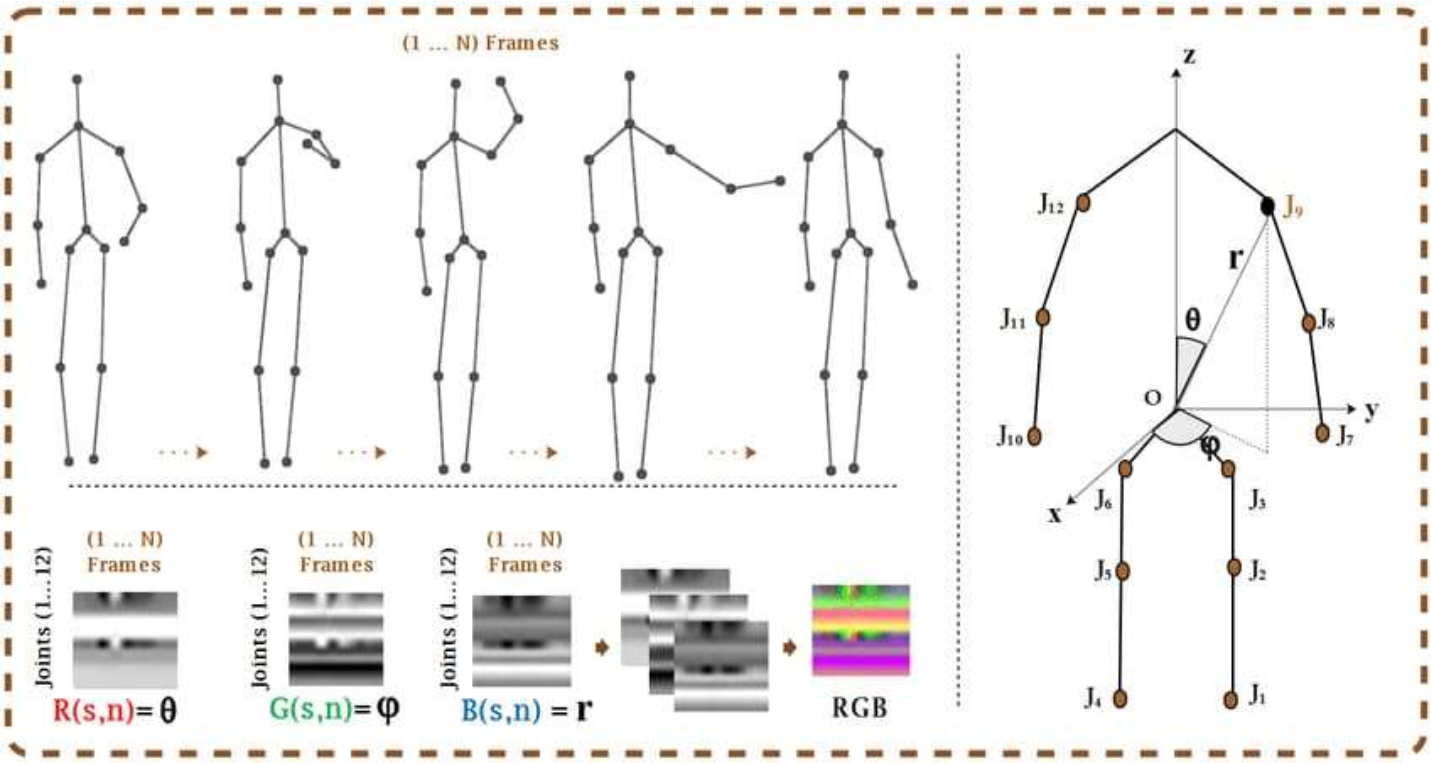

Fig. 2. Moving Joints Descriptor (MJD).

Moving Joints Descriptor (MJD) can be explained, with the help of a similar example of Human action of drawing in circular motion from the above represented dataset, left-top: Skeleton sequence, firstly: Creation of RGB Moving Joints Descriptor picture, secondly, Skeleton dummy model represents the 3 3D circular coordinates of joint j 9. Where N: entire range of frames, s: number of joints, and n being variety in the body.

Year	Title	Techniques used	Classifier Used	Data Set	Accuracy
2016	Action Recognition Based on Joint Trajectory Maps Using Convolutional Neural Networks	Using ConvNets in three orthogonal planes from the skeleton sequences.	Late fusion of the three ConvNets	(MSRC-12), G3D dataset and UTD multimodal hu man action dataset (UTD-MHAD)	94.85%
2017	A New Representation of Skeleton Sequences for 3D Action Recognition	starts by generating clips of skeleton sequences then transforms it into three clips each consisting of several gray Images then passing through MTLN.	Multi-Task Learning Network	NYU RGB+D	83.52%
2016	Skeleton Based Action Recognition with Convolutional Neural Network	we first detail how to represent a sequence as an image, and then we introduce our hierarchical model for skeleton based action recognition	None	ChaLearn ges ture recognition, Berkeley MHAD	91.21%
2016	Combining Multiple Sources of Knowledge in Deep CNNs for Action Recognition	we present two ways of combining spatial and temporal cues in deep architectures for action recogni tion	magnitude of optical flow	UCF 101 HMDB 51	89.1%
2015	UTD-MHAD: A Multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor	Action Recognition Using Depth And Inertial Sensor Fusion	CRC classifier	UTD-MHAD	67.2%
2015	Modeling Transition Patterns Between Events for Temporal Human Action Segmentation and Classifification	temporal segmentation and classifification method that accounts for transition patterns between events of interest	Support Vector Machine	a public CMU-MAD benchmark dataset,smart room dataset	74.40%
2020	Human Action Recognition Based on Improved Fusion Attention CNN and RNN	Fusion KeyLess Attention combining with the forward and backward bidirectional LSTM	None	HMDB51 and Hollywood2	59.6%
2019	Study on 3D Action Recognition Based on Deep Neural Network	proposes to use depth-map images from 3D cameras to supplement the 2D surveillance camera systems to detect objects and monitor abnormal events in detail	Depth Motion Map (DMM)	KETI dataset NTU RGB+D KETI RGB+D	61.5%
2020	A Cuboid CNN Model with an Attention Mechanism for Skeleton-based Action Recognition	cuboid CNN model with an attention mechanism	the cuboid feature arranging	CAS-YNU MHAD, NTU RGB+D dataset,UTD-MHAD dataset,UTKinect-Action3D dataset	96.10%

Table. 1. Literature Survey

II. LITERATURE REVIEW

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

In [5], a strategy is proposed which is minimal and viable nonetheless simple technique to code spatiotemporal information conveyed in 3D skeleton successions into various second pictures, raised as Joint mechanical wonder Maps (JTM), and ConvNets are embraced to utilize the discriminative alternatives for ongoing activity acknowledgment. The proposed approach has been assessed on 3 public benchmarks, i.e., diverse public datasets accessible on the web and accomplished the cutting edge results. [6] The proposed strategy has been assessed on 3 public benchmarks, i.e., diverse public datasets accessible on the web and accomplished the best in class results. The proposed strategy initially changes every skeleton arrangement into three clasps each comprising of a few edges for spatial worldly component getting the hang of utilizing profound neural organizations. Each clasp is produced from one channel of the barrel shaped directions of the skeleton succession. Each packaging of the made slices addresses the transient information of the entire skeleton progression, and combines one explicit spatial association between the joints. The entire fastens consolidate different edges with different spatial associations, which give significant spatial essential information of the human skeleton. We propose to use significant convolutional neural associations to learn long stretch transient information of the skeleton course of action from the housings of the delivered fastens, and a while later use a Multi-Task Learning Network (MTLN) to together deal with all edges of the made slices in relating to combine spatial hidden information for movement affirmation. Exploratory results evidently show the practicality of the proposed new depiction and feature learning procedure for 3D action affirmation.

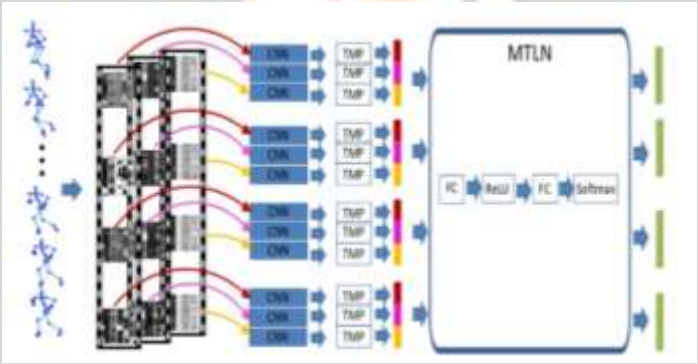

Fig. 3. Existing MTLN model.

[4] This strategy proposes partner start to finish gradable engineering for skeleton principally based activity acknowledgment with CNN. From the start, we may show the skeleton arrangement as a lattice by joining the joint directions in each occasions and putting together these vector pictures during the composed record request. At that point the removed network is streamlines into a picture and is made ordinary to utilize the length of the variable to be a misfortune. The last picture is transferred into a CNN model for highlight extraction and acknowledgment. For explicit design of each pictures, the straightforward max-pooling assumes an indispensable part on deliberation highlight determination, alongside fleeting recurrence change, this may get a great deal of separated joint information for different activities alongside, tending to the variable-recurrence issue. [8] This paper puts forward, approaches for consolidating various wellsprings of information in profound learning. To start with, we propose highlight intensification, where we utilize an assistant, hand-made, highlight (for example optical stream) to perform spatially shifting smooth gating of middle 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 understanding. We test these techniques in the setting of activity acknowledgment where data from spatial and transient signs is helpful, getting results that are practically identical with cutting edge techniques and outflank strategies utilizing just CNNs and optical stream highlights. [9] This paper portrays an uninhibitedly available dataset, named UTD-MHAD, which contains four momentarily synchronized data modalities. These modalities join RGB accounts, significance

chronicles, skeleton positions, and inertial signs from a Kinect camera and a wearable inertial sensor for an extensive arrangement of 27 human exercises. Test outcomes are given to show how this data base can be used to consider mix pushes toward that incorporate using both significance camera data and inertial sensor data. This public area dataset is useful for multimodality research practices being coordinated for human action affirmation by various assessment get-togethers. [10] proposed a fleeting division and arrangement procedure that addresses progress plans between events of interest. We apply this method to normally perceive striking human movement events from chronicles. A discriminative classifier is used to see human action events and a capable exceptional programming computation is used to commonly choose the start and completing transient pieces of apparent human exercises. The basic differentiation from past work is that we present the showing of two kinds of event progress information, explicitly event change areas, which get the occasion de-signs between two consecutive events of interest, and event progress probabilities, which model the advancement probability between the two events. Test outcomes show that our strategy out and out improves the division and affirmation execution for the two datasets we attempted, in which specific advancement plans between events exist. [18] This paper presents of significance sensors, for instance, Microsoft Kinect have driven investigation in human action affirmation. Human skeletal data accumulated from significance sensors pass on an immense proportion of information for movement affirmation. While there has been amazing progression, in actuality, affirmation, by and large existing skeleton-based procedures ignore the way that not all human body parts move during various exercises, and they disregard to think about the ordinal spots of body joints. Here, and prodded by how an action's class is constrained by neighborhood joint turns of events, we propose a cuboid model for skeleton-based action affirmation. Specifically, a cuboid getting sorted out framework is made to arrange the pairwise migrations between all body joints to get a cuboid movement depiction. Such a depiction is throughout coordinated and allows significant CNN models to fixate examinations on exercises.

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

1. Regardless of the way that CNN is outfitted with extraordinary component extraction and arrangement in a considerable lot of the PC vision issues, the CNN model can't characterize the activities accurately explicitly when the info pictures don't furnish with discriminative highlights.

2. The current models all characterize the activities dependent on the worldwide spatial and transient data found in the skeleton successions. This requires the commotion circulation in various portions of a similar grouping to be reliable. Subsequently prompting the acknowledgment rate been chopped down, if the information blunder of neighborhood sections in the info groupings is featured.

3. It is realized that every one of the organization meets to an alternate neighborhood minimum, regardless of whether each organization were to be prepared on a similar information methodology. The exhibition increments when joining various organizations along with straightforward late combination approach because of every nearby minima having a marginally unique information. Consequently, there is a need to multiplicatively consolidate numerous CNNs to arrange the pictures.

4. An end segment isn't needed. Despite the fact that an end may audit the central matters of the paper, don't duplicate the theoretical as the end. An end may expand on the significance of the work or propose applications and augmentations.

IV. CONCLUSION

Human action acknowledgment is indispensable for various PC vision applications that demand information of people lead, including reconnaissance for public security, human-c cooperation applications and mechanical innovation. Regardless, action affirmation in concealed pictures is trying assignment on account 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 surveyed on various existing technologies, which have proven to be of a maximum accuracy of 97%. The most efficient of all being considered to be the usage of CNN’s for action recognition, though there are a few drawbacks even in this method since discriminative images are not provided during the input.

V. 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".