https://doi.org/ 10.33472/AFJBS.6.9.2024.2092-2102

AFJBS

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper

Open Access

Human Pose Recognition using Deep Learning

U. Chaitanya[1], Teenderu Vaishnavi[2], Anuradha Kumari[3], Veerla Vikas[4]

Assistant Professor[1], UG Students[2,3,4]

Dept of Information Technology, Mahatma Gandhi Institute of Technology

.

Article History

Volume 6,Issue 9, 2024

Received: 26-03-2024

Accepted : 30-04-2024

doi: 10.33472/AFJBS.6.9.2024.2092-2102

Abstract: Human Pose Recognition (HPR), the work of estimating the spatial configuration of human's body through a video has shown notable recognition in deep learning and artificial intelligence communities. HPR is a demanding technology that is used across numerous applications to estimate human pose actions in real-time.HPR is the process of identifying the activities being performed by human through analysing video data traces of their movements. Our proposed methodology uses real world dataset to draw precise labels with the probabilities for corresponding human poses. The primary aim of our proposed methodology is to label activities by analysing video data. This is achieved initially by taking real-time video input from the user using OpenCV and performs pose recognition. In our proposed methodology, we used six different categories such as sitting, standing, walking, pain, falling, sleeping using a Mediapipe 32-point landmark key-point detection model and hybrid CNN-LSTM model. The CNN-LSTM enhances pose recognition by effectively capturing both spatial and temporal features. However, probabilities of poses such as sitting, standing, walking, pain, falling, sleeping are analysed and achieved overall 96.5% accuracy.

Index terms – *Human Pose, Mediapipe, convolutional neural networks (CNN), Long short-term memory networks (LSTM), Hybrid CNN-LSTM model, OpenCV*

1. INTRODUCTION

Human Pose Recognition aims to recognize the activity by processing the video frame. We propose a framework that uses Mediapipe, CNNs and LSTMs networks for correct and accurate human pose recognition.

First, we are using MediaPipe to extract key landmarks that represents human body joints in input video frames. These key landmarks are used as input data for following processing stages.

After using Mediapipe , we are using CNN architecture to extract spatial features from the detected key points, that helps in identifying crucial spatial patterns and relationships among body landmarks.

Afterwards, we provide the extracted spatial features into an LSTM model to network temporal features over consecutive dependencies frames. The LSTM rigorously obtains temporal dynamic features in sequences of human poses, this helps in identifying complex motion patterns and gestures.

Our proposed framework is trained on big-scale human actions datasets that is normal and abnormal activities, that allows to gain insights on features and temporal dynamics straight from data. Experimental result output display the efficiency of our methodology, achieving continuity performance on various human pose recognition environment.

Overall, our proposed methodology gives a comprehensive result for human pose recognition, with the help of MediaPipe, CNNs, and LSTMs to achieve greater performance in displaying both spatial and temporal aspects of human motion.

2. LITERATURE SURVEY

Lamiyah Khattar et al. [1] proposed a 2-D CNN model for analysing and monitoring human poses. CNN top in image processing by making use of convolutional matrices to carry out functions like Finding pattern, blurring and sharpening images. Whereas, LSTM network, model of Recurrent Neural Networks (RNNs), are good at refining sequential data, that helps them in time series analysis and classification. Period, both 2-D CNN ad LSTM excel same acceleration changes, their methodology differs. LSTM layers are enhanced for networking sequence data, while CNNs coordinated in tracing spatial relations among data.

Amy Bearman et al. [2] focused on CNN model for image identification tasks. This CNN consists of five layers and 3 layers that are inter connected fully such as max pooling, non-linearities, and normalization layer. The end layer executes the SoftMax function for prediction. By exploring CNN usage to regression task, they applied CNNS for key point detection straight from input images which concentrate on interaction among parts of body. Although, Deepness of Regression CNN was bounded by GPUs memory restrictions.

Shreyank N Gowda et al. [3], [4], [5] concentrated on enhancing activity recognition precision by refined frame detection, especially sharping edited, trimmed and short input videos. However, it introduces a joint approach that considers constructive distribution of frames all over the video that collects action sequence. This strategy carries out frame detection methods across benchmarks such as FCVID, UCF101 and ActivityNet. This helps in enhancing versatility across various datasets in action recognition operations.

Valentin Bazarevsky et al., [6], [7], [8] proposed a Blaze Pose model that is a light weight architecture for human pose estimation that is utilized for hands on smart phones. This also locates 33 key points on body on a single human in very less time by producing over 30 frames. Their innovation contributes a unique innovation for body activity tracking and a neural layer structure for activity recognition. This helps in improving model's efficiency and make sure to estimate pose in real-time by making it ideal for applications such as sign language recognition, Human activity recognition and gym tracking.

Jupalle Hruthika et al. [5]introduced a two-branch network that includes two stages. First, pre-train the model utilizing Lpre-train loss with Adam optimizer and 0.001 learning rate. Following a LT loss with a decreasing learning rate of 0.1. During the assess, 2D-to-3D branch identifies relative activities for robustness. The contribution of proposed work includes tracing temporal features in videos that contributes geometric conversion from one frame to another by eliminating usage of computations in optical flow. This helps in gaining performance by integrating spatial techniques for further work.

Mohamed S. Abdallah et al.[11] proposed an idea to generate real-time DSLR for sign language recognition system to remove the communication gap among hearing-impaired people and the general people. The methodology they followed are, implementing deep neural networks, utilizing the GRU and the 1DCNN models. later, MediaPipe framework are integrated with these models for efficient extraction and processing of poses from video data and detecting hand gestures. The obtained solution was verified on a new United States of America gesture Language dataset (DSL-46) and other benchmark datasets, obtaining high accuracy rates of 98.8% with DSL-46, 99.84% with LSA64, and 88.40% with LIBRAS-BSL. This methodology focuses on enable ling fast and exact DSL recognition.

sahak kaghyan et al. [9], 10] investigates the effectiveness of the K-Nearest Neighbor (K-NN) algorithm in classifying human activities using smartphone accelerometer data. Two experiments are done that is one for data gathering on Android smartphones using the accelerometer, and second is for activity classification using K-NN on a desktop platform. The K-NN algorithm depends on the maximum vote of the K nearest neighbors in feature space. But K-NN has limitations e.g sensible to noisy features and dependency on training data. Future scopes include exploring the enhancement to K-NN, allowing to take data from various sources like GPS, and finding alternative classification methods like decision trees. Results shows that accelerometer data facilitates exact activity recognition, but activities limited to hand or mouth movements pose challenges.

Iveta Dirgova Luptakova et al. [12] explores the adaptation of transformer models, designed for NLP and vision tasks, to analyze motion signals for human activity recognition. It underlines the significance of accelerometer and gyroscope data from smartphones in sports, healthcare system, and human-robot interaction. The adapted transformer model utilizes its self-attention mechanism to capture dependencies within sequence data, achieving a remarkable average prediction accuracy of 99.24% compared to traditional methods 89.67% on a large smartphone dataset. The architecture of the transformer network includes multi layers, full layers, normalization layers, dropout layers, and residual connections, enabling efficient capture of connection between features within time steps. Additionally, the paper talks the transformer's uses in image categorization tasks, where it changes convolutional layers by dividing images into patches and implementing attention mechanisms. Though the transformer yields scalability and speed advantage due to large set of image dataset.

S. M. Salahuddin Morsalin et al. [14], [20] proposed a unique method to single human pose estimation, specially focusing on elderly monitoring, fitness activities and mobile applications. It demonstrates a framework that involves of a initial network to start pose estimation and an IFC network to refine. The IFC network targets on high-level limitations on global pose correction and body surface rectification, decreasing the impact of body joints and body movement part. This method includes basic network pose estimation with use of MediaPipe tools, followed by IFC network processing with a determined loss function. Front and back process processes focus on optimization of the network parameters based on the loss function. Regarding human pose estimation, the proposed work can be extended by inter connecting additional methods or classifications on higher stage of the pose estimation results. This helps in the estimation of specific abnormal poses such as yoga poses or body lifting actions by training machine learning on labeled data using the Mediapipe pose as input frame features. The framework's superior accuracy and robustness make it fit for fitness and tracking applications.

Daniel Wagner et al. [13] deployed the potential of identifying human activity using a 2-D Convolutional Neural Network (CNN) and a Support Vector Machine (SVM). The main concept includes transforming a one-dimensional network to a two-dimensional form, particularly illustrated through the creation of pictures.

Subsequently, it is focused on a pretrained deep neural network, such as AlexNet, precisely taking features by eliminating being particular trained on such data. For classification, an SVM is used with combination of linear and nonlinear kernels, that shows classification performance across various datasets. Moreover, the large dataset consists two classes, that yields desired outcomes. particularly for both datasets, it is important that the linear kernel gives results similar to the nonlinear one. Based on these outputs, further enhancement of the proposed methodology for extraction of features using a 2-D CNN is justified, correctly with the construction of a customized neural network architecture comprising a classification layer.

3. METHODOLOGY

3.1 Proposed Architecture

HPR uses CNN-LSTM hybrid model architecture to manage the time series data. Contrasting to LSTM's dimensions that include 3-dimensional data [samples, time steps, features], CNN-LSTMs takes 4-dimensional data with dimensions [samples, nsteps, nlength, features].

Sample denotes number of instances in the dataset, that corresponds to an image. nsteps defines the various number of steps that divide time sequence into segments. nlength refers to length in specific timestep after performing division, it basically represents frames numbers in all of the segment. Feature represents information about the positions of keypoint of the body. Thus, various layers included in our model are Time Distributed Conv1D layer, Time Distributed Conv1D Droupout layer, Time Distributed MaxPooling 1D layer, Time Distributed Flatten layer, LSTM Dropout Layer that enables optimizing sequential data with spatial domains by convolutional networks and temporal domains captured using LSTM networks as described in Fig. 1

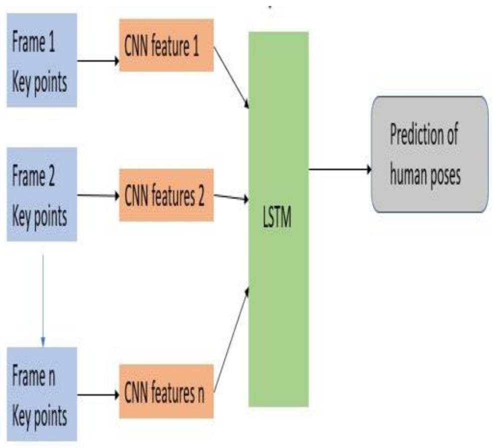
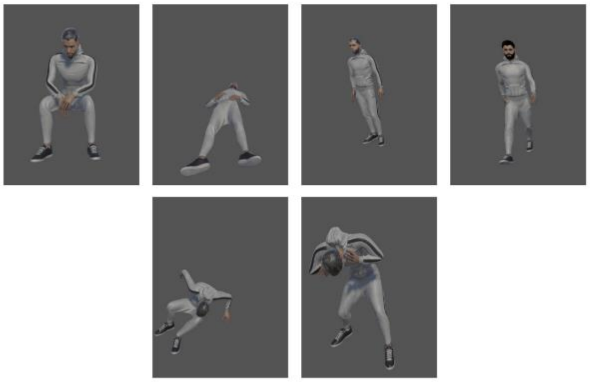


Fig.1 Proposed HPR Architecture

3.2 Dataset

Instead of collecting real-time data through a rehabilitation centre, we generated data from Unity3D that allows to simulate numerous activities and typical scenarios. We took 52,128 samples for training the model. Our data is classified into two groups as displayed in Fig 2 [11]. First includes normal activities that encompasses daily poses such as sitting, walking, sleeping and standing. Second group consists of abnormal activities that indicate poses different from normal and requires attention such as falling, coughing and experiencing pain. By training both normal and abnormal activities to



our model, we distinguish between various activities.

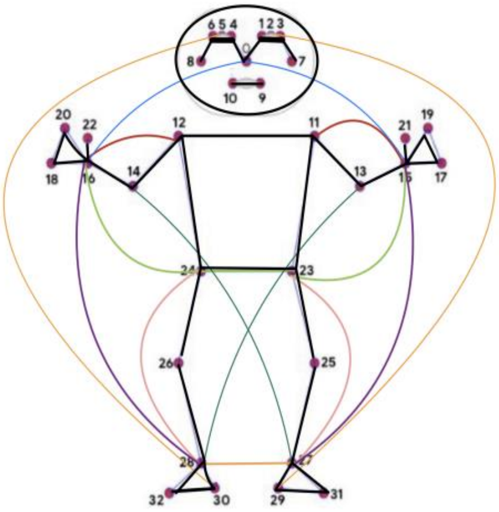


Fig.2 Unity 3D Dataset [11]

15-16	15-11	12-16	15-23	16-24
23-24	16-28	15-27	23-27	24-28
28-27	5-30	2-29	14-27	13-28
0-15	0-16	26-25	31-32	

Fig.3 Human Skeleton Key points [12]

3.3 Data Pre-processing

To develop our simulations for key point location, we divided the data into single frames and selected key points using estimated distances and landmarks for different activities. As a result, we gained both the original 99 key points and extra 19 features which calculate distances between key point landmarks. This feature set was normalized into our model, to leverage all classification to enhance performance of our pose estimation system as displayed in Fig 3 [12]

3.4 Proposed work

We utilized the MediaPipe library to locate 33 key points on human body poses. Each key point briefs us with three-dimensional coordinates (x, y, z), deriving in a sum of 99 features. We additionally computed 19 features that gives distances between specifically chosen pairs of key points. The chosen pairs were picked based on the sub-key point distance are mathematically calculated utilizing Euclidean distance (Ed) given in Eqn. (1) and selected positions among the various activities. Thus, resulting in gaining a total of 118 feature points that serve as landmark locations as input for using in model in future.

$$\mathbf{Ed} = \sqrt{(\boldsymbol{keypointx_i} - \boldsymbol{keypointy_i})^{\mathbf{2}}} \quad (1)$$

Leveraging CNN model carries out two stages of convolutional layers such as ReLU activations and max pooling. ReLU convolutional layer mainly focuses on input channel transformation to 32 feature maps and sets all negative values to zeros and leaves positive values unchanged to harness non-linearity, and the max pooling layer increases feature maps to 64 thereby decreasing computational complexity and monitor overfitting. These CNN networks are categorized to gain spatial features from input channel [20].

LSTM layer monitors the time series data. It takes results from CNN that has been reshaped for fitting LSTM's requirements and to extract temporal features. The LSTM is allowed to be bidirectional such that it learn from past and future data. LSTM generally includes 3-dimensional data with dimensions [samples, time steps, features].

Fully connected layer computes the LSTM's results to the various classes for categorization of activity. A hybrid CNN-LSTMs expect 4-dimensional data [samples, nsteps, nlength, features].

3.5 EXPERIMENTAL RESULTS

After running 120 epochs, our model accomplished in achieving accuracy up to 95%. In this iterative process, the loss continuously decreased, sparking a consistency in the model's efficiency and labelled activities as shown in Fig. ,5,6,7,8 and 9.

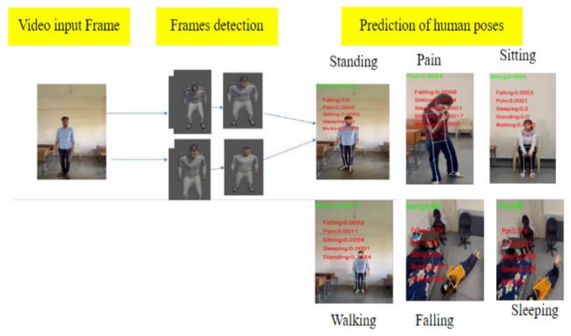


Fig.4 Proposed HPR Results

TABLE I. Activities Sample ratio

Activity	Class Label	Training Samples	Video Frame
	Sitting	1596 (18.31%)	326 (19.8%)
	Standing	1202 (13.8%)	510 (31.1%)
	Walking	2051 (23.5%)	520 (31.7%)
	Sleeping	1498 (17.2%)	39 (2.38%)
	Falling	1017 (11.67%)	115 (7.01%)
	Pain	1349 (15.48%)	129 (7.87%)

Table.1 depicts total of 8713 sample dataset frames are utilized for training the model, While to test the model, 1639 sample of data are used.

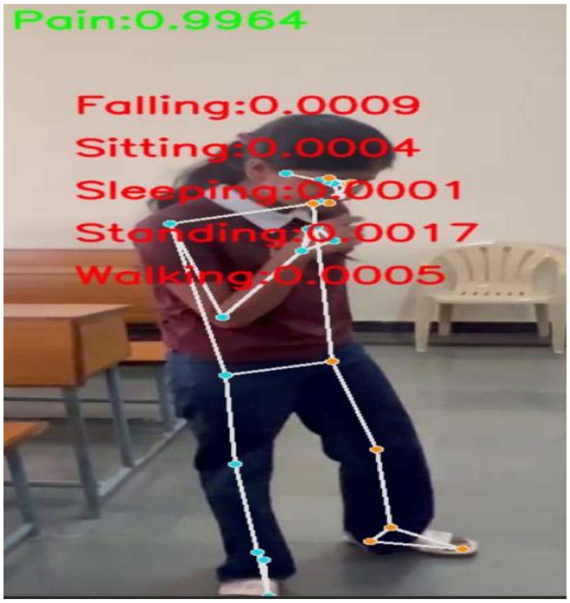


Fig.5 Pain Pose

Fig.5 depicts the landmarks on the human body that predicts pose as pain with corresponding probability of 99.64%.

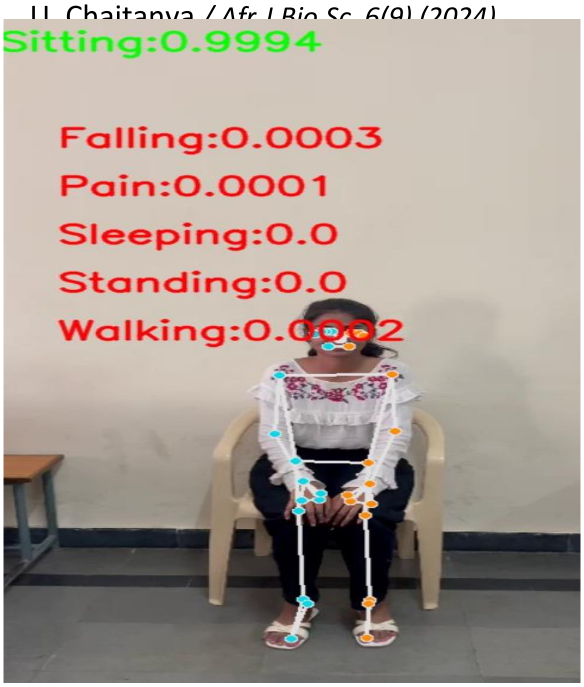


Fig.6 Sitting Pose

Fig.6 depicts the landmarks on the human body that predicts pose as sitting with corresponding probability of 99.94%.

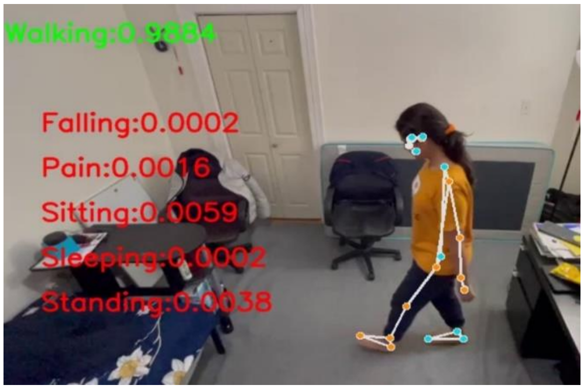


Fig.7 Walking Pose

Fig.7 depicts the landmarks on the human body that predicts pose as walking with corresponding probability of 98%.

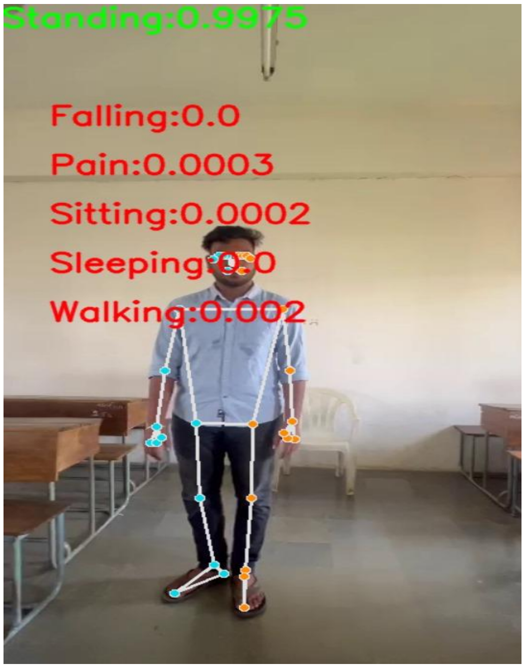


Fig.8 Standing Pose

Fig.8 depicts the landmarks on the human body that predicts pose as Standing with corresponding probability of 99.75%.

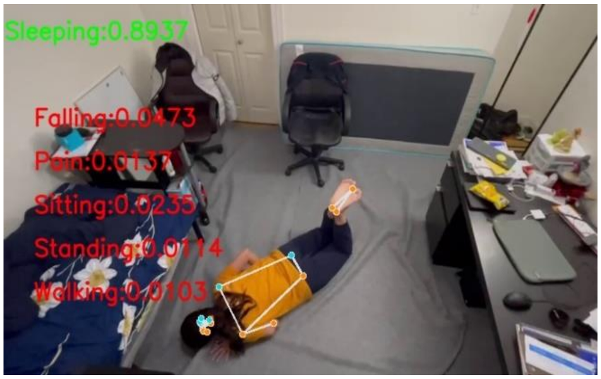


Fig.9 Sleeping Pose

Fig. 9 depicts the landmarks on the human body that predicts pose as sleeping with corresponding probability of 89.17%.

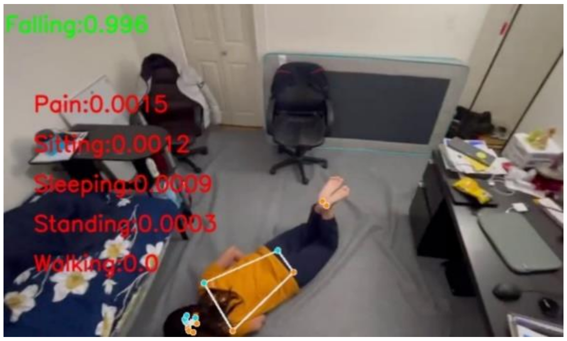


Fig.10 Falling Pose

Fig. 10 depicts the landmarks on the human body that predicts pose as falling with corresponding probability of 99.6%.

TABLE II. Comparison of Accuracy (%)

Method	Proposed	[13]	[14]	[15]	[16]
Walking	**99.40**	95.61	99.19	97.78	84.84
Sitting	**99.94**	92.96	87.98	96.77	79.46
Standing	**99.75**	96.43	97.37	87.08	80.69
Total	**99.69**	95	94.84	93.87	81.66

After performing a comparative analysis of our proposed methodology (CNN-LSTM) and existing methodology (CNN) for poses such as sitting, standing, walking, pain, falling, sleeping. the proposed methodology improves its performance by increasing 1.5% accuracy.

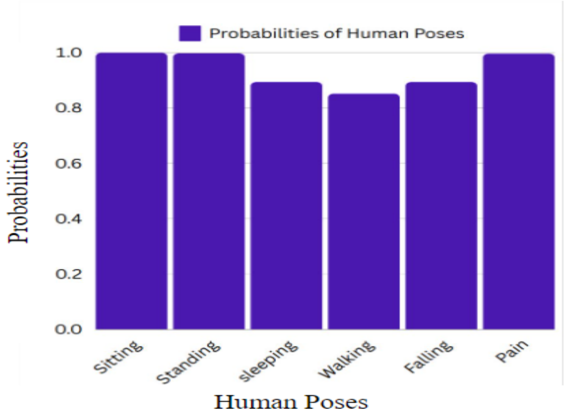


Fig.11 Bar chart depicting poses with corresponding probabilities

CONCLUSION

The proposed HPR using CNN-LSTM and Mediapipe deploys an innovative solution of various applications like assisting physically disabled people, monitoring elderly people, sports tracking, Fitness tracking. This improvement depicts the effectiveness of our approach in accurately predicting various poses. Our proposed methodology displaying probabilities for various poses such as sitting, standing, walking, falling, pain, and sleeping with corresponding probabilities 99.64%, 99.94%, 99.75%, 85.17%, 99.6%, and 85.17%. Hybrid CNN-LSTM model with the combination of mediapipe actively integrates all the layers and achieved accuracy of 96.5% and improved 1.5% accuracy compared to the existing CNN model.

Proposed methodology can be extended on data augmentation methods to leverage training data and plan on adding more poses such as finding poses between sitting and standing, bending, actions, construct the trained model for real-world usage by performing rigid testing and calculating its efficiency for hands on utility, adding voice feature makes feasible for assisting that makes job easier.

REFERENCES

1. Lamiyah Khattar, “Analysis of Human Activity Recognition using Deep Learning,” 11th International Conference on Cloud Computing, Data science & Engineering, pp. 100-104, 2021.
2. C. D. Amy Bearman, “Human Pose Estimation and Activity Classification Using Convolutional Neural Networks,” pp. 1-8, 2018.
3. Madapuri, R.K., Mahesh, P.C.S. HBS-CRA: scaling impact of change request towards fault proneness: defining a heuristic and biases scale (HBS) of change request artifacts (CRA). Cluster Comput **22** (Suppl 5), 11591–11599 (2019). https://doi.org/10.1007/s10586-017-1424-0
4. Swetha, A. ., M. S. . Lakshmi, and M. R. . Kumar. “Chronic Kidney Disease Diagnostic Approaches Using Efficient Artificial Intelligence Methods”. International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 1s, Oct. 2022, pp. 254
5. Thulasi , M. S. ., B. . Sowjanya, K. . Sreenivasulu, and M. R. . Kumar. “Knowledge Attitude and Practices of Dental Students and Dental Practitioners Towards Artificial Intelligence”. International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 1s, Oct. 2022, pp. 248-53.
6. Rudra Kumar, M., Gunjan, V.K. (2022). Machine Learning Based Solutions for Human Resource Systems Management. In: Kumar, A., Mozar, S. (eds) ICCCE 2021. Lecture Notes in Electrical Engineering, vol 828. Springer, Singapore. https://doi.org/10.1007/978-981-16-7985-8_129
7. Shreyank N Gowda, “SMART Frame Selection for Action Recognition,” Computer Vision and Pattern Recognition (cs.CV), no. Available: https://arxiv.org/abs/2012.10671, 2020.
8. G. Valentin Bazarevsky, “BlazePose: On-device Real-time Body Pose tracking,” Computer Vision and Pattern Recognition (cs.CV), no. [online] Available: https://arxiv.org/abs, 2020.
9. Jupalle Hruthika, “DEEP LEARNING BASED HUMAN POSE ESTIMATION USING OPENCV,” IJIERT, vol. 7, pp. 246-253, 2020.
10. Mohamed S. Abdallah, “Light-Weight Deep Learning Techniques with Advanced Processing for Real-Time Hand,” IJIMA, vol. 1, pp. 146-154, 2012.
11. Sahak Kaghyan, “ACTIVITY RECOGNITION USING K-NEAREST NEIGHBOR ALGORITHM ON

SMARTPHONE WITH TRI-AXIAL ACCELEROMETER," IJIMA, pp. 146-154, 2012.

12. M. K. Iveta Dirgová Luptáková, "Wearable Sensor-Based Human Activity Recognition with Transformer Model," p. vol: 22, 2022.
13. C. Ming-Hwa Sheu, "Improvement of Human Pose Estimation and Processing with the Intensive Feature Consistency Network," IEEE, vol. Vol: 11, 2022.
14. K. Daniel Wagner, "Activity Recognition using Inertial Sensors and a 2-D Convolutional Neural Network," IEEE, 2017.
15. Unity 3D, https://unity.com/.
16. Shivam Nikam, "HUMAN ACTIVITY RECOGNITION USING OPENCV AND GOOGLE MEDIAPIPE," IRJMETS, vol. 5, 2023.
17. Niloy Sikder, Human Activity Recognition Using Multichannel Convolutional Neural Network, 2019.
18. L. O. Davide Anguita1, Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support, 2013.
19. Bhattacharjee, "A Comparative Study of Supervised Learning Techniques for Human Activity Monitoring Using Smart Sensors," Second International Conference on Advances in Electronics, Computers and Communications (ICAECC), pp. 1-4, 2018.
20. Rudra Kumar, M., Gunjan, V.K. (2022). Peer Level Credit Rating: An Extended Plugin for Credit Scoring Framework. In: Kumar, A., Mozar, S. (eds) ICCCE 2021. Lecture Notes in Electrical Engineering, vol 828. Springer, Singapore. https://doi.org/10.1007/978-981-16-7985-8_128
21. G. P. Bota, "A Semi-Automatic Annotation Approach for Human Activity," Sensors (Basel), vol. vol.19, pp. 1-23, 2019.
22. Almaslukh, "An Effective Deep Autoencoder Approach for Online Smartphone Based Human Activity Recognition," Int. J. Comput. Sci. Netw. Secur, vol. vol 17, pp. 160-165, 2017.
23. Chaitanya Yeole, "Deep Neural Network Approachesfor Video Based Human Activity Recognition," IJISRT, vol. 6, 2021.
24. Jianfeng Zhang a, "Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas," ScienceDirec, vol. 561, 2018.