Human Action Recognition Using LRCN & LSTM

1st Siddharth Bhorge
Electronic And Telecommunication
Vishwakarma Institute Of Technology)
Pune , Indian
siddharth.bhorge@vit.edu

2nd Swapnil Patil
Electronic And Telecommunication
Vishwakarma Institute Of Technology)
Pune , Indian
swapnil.patil21@vit.edu

3rd Devyani Ushir
Electronic And Telecommunication
Vishwakarma Institute Of Technology)
Pune , Indian
devyani.ushir21@vit.edu

4th Komal Shinde
Electronic And Telecommunication
Vishwakarma Institute Of Technology)
Pune , Indian
komal.shinde21@vit.edu

5th Aditya Vhanmane
Electronic And Telecommunication
Vishwakarma Institute Of Technology)
Pune , Indian
aditya.vhanmane21@vit.edu

***Abstract*—Recent developments in artificial intelligence have enabled the world to detect objects, learn their surroundings, and forecast the next sequences. The cost of surveillance systems is reduced as a result of the development of embedded technology. The surroundings are being captured by the surveillance equipment and are being kept in memory. To interpret the environmental data we collected and understand the scenario, deep learning is used. This Paper examines the notion of using film to identify human action and behavior. Additionally, this Paper suggests combining LSTM and CNN for analyzing the video. Convolution processing transforms the input into relevant spatial information. To create temporal features, the collected features are fed into lengthy short- term modules and Long term recurrent convolution network. The hypothesized attention elements were fed by the feature maps of the LSTM and LRCN. It captures in the video's frame the really valuable instructive aspects. Using video, these models can identify human behaviors. The experimental findings demonstrated that the proposed model performed more accurately and efficiently.**

Keywords— Human Activity, Convolutional neural networks, long short term memory, frame extraction, Recognition, LRCN.

I. INTRODUCTION

Due to its use in numerous industries and the growing need for home automation and convenience offerings for the elderly, human pastime attention has been a popular examine subject matter in the latest years . To enhance the first-rate existence of residents inside the domestic environment, activity reputation in a clever domestic using a straightforward and all-cause sensor and deep gaining knowledge of is any such this is attracting plenty of attention . With the use of deep gaining knowledge of, primary and complex sports in actual-international situations are to be recognized and recognized. In interactions between humans and computer systems, the reputation of human conduct is essential. In our task, we rent Convolution Neural Networks (CNNs), which excel at managing picture statistics, and long brief period reminiscence Networks (LSTMs), which excel at coping with collection information. but, by combining the 2, you can achieve the quality of both worlds and solve difficult troubles categorization. In this look, we superior a convolution neural network together with a long term memory network to recognize human hobbies on video. We employed two wonderful TensorFlow architectures and strategies to do that. This test made use of the UCF50 data set.

II. LITERATURE SURVEY

Huang et al. [1] proposed a method for recognizing human movements in video sequences of varying resolutions using the Hierarchical Filtered motion model and the nearest Neighbor classifier with HOG feature. The method involved combining datasets from the KTH dataset, which contains six precise human movements (strolling, walking, taking walks, boxing, hand-clapping, and elevating of the hands) and was used for training, with the MSR dataset, which contains hand-clapping, boxing, and hand-waving, and was used for testing. Their study demonstrated the effectiveness of their proposed method in accurately recognizing human movements.

Lu and Nguyen et al. [2] in this research ,they use cutting-edge deep learning techniques to research the trouble of human behaviour recognition. with a view to obtain sufficient recognition accuracy, both spatial and temporal records had been acquired. Deep mastering models could be prolonged with the aid of growing the intensity or width of network layers to enhance method accuracy. Convolutional deep learning improves performance with out increasing complexity. within the future, we can contain the attention module and spectrum statistics into the proposed models to improve accuracy.

Nguyen et al. [3] gives a spatial-temporal interest-aware feature pooling method that is evaluated on 3 properly-regarded video action statistics units. especially, ninety five.3% on UCF sports (better via 4.0%), 87.9% on YouTube (better through 2.5%), and comparable outcomes on Hollywood2 (data set). The features of the system include the merging of visual attention with SPM-based pooling techniques. This approach is a highly successful method for identifying actions in a range of realistic videos. As a result, the suggested STAP can produce state-of-the-art results on a variety of widely used action recognition data sets.

Yuan et al. [4] an effective method for identifying human motion has been given in this research. For this gadget to file human movements, simply one standard digital camera is needed. In numerous conditions, the gadget can efficaciously extract a contour and do away with shadows. even though the system has a high rate of reputation, it nevertheless has sizable boundaries. They take the frame series's human outlines and extract them using a -layered heritage that is based totally on each chromaticity and gradient. The reason for the vision-based human interest popularity is to provide a trustworthy, natural technique to perform this. nothing desires to be affixed to the frame with this technique, not like sensor or marker structures. The purpose of the estimation technique is to become aware of the movement that is maximum likely to occur given the parameters.

Reddy et al. [5] in this Paper, movement popularity in sizeable categories of unrestricted online films is discussed as an exceedingly hard subject matter. We propose leveraging motion features and scene context information from transferring and desk-bound pixels within the keyframes to tackle the difficulty of a massive (50 movements)dataset. On datasets like UCF11 (87.19%), UCF50 (76.90%), and HMDB51 (27.02%), the counseled method plays first class. We verified that the movement descriptors lose discriminative electricity because the number of classes rises.

Shuiwang Ji et al. [6] 3D CNN (Convolutional Neural Networks) models for motion recognition are used in this paper. through executing 3D convolutions, these fashions create functions from each spatial and temporal measurement. it is vital to lay out a deep structure that produces numerous channels of facts from nearby input frames to perform convolution and sub-sampling for my part in every channel. information from all channels is used to create the very last characteristic depiction.

Leung et al. [7] this paper focus on the reason that the articulated motion facts are tremendously dimensional in both the spatial and temporal domain names, recognizing human moves is a hard procedure. separating the human body into awesome frame components, as indicated through the placements of the human skeletal joints, and appearing popular based totally on those part-based element descriptors are powerful methods to cope with this unpredictability.

Dawar et al. [8] in this paper, deep learning-primarily based sensing is offered. Fusion machine to locate and pick out essential moves in streams of ongoing action. each sensing technique segments all moves earlier than identifying those which can apply to a sure utility. two applications of the proposed machine are checked out: one involves transition motions for domestic healthcare monitoring and the opposite involves hand gestures for smart TVs. The device makes use of inertial statistics from a wearable inertial sensor and depth pictures from an intensity digicam. the application of choice-level fusion to the movements of interest detected via both modalities.

Zhou et al. [9] this paper suggests a technique for concurrently locating the vital spatial and temporal aspects of interest in instructional videos. The technique represents activities by using dense trajectories that had been taken from movies as local characteristics. We provide a set of rules that segments a video and numerous distinct foreground moving objects through the usage of a trajectory cut up-and-merge algorithm. To make segmentation simpler, the natural temporal smoothness of human motions is taken gain of. Then, we infer the spatial and temporal extents of the applicable activity through the use of the latent SVM framework on segmentation findings.

Xia and Chen et al. [10] the approach for identifying human actions as time series of recognizable 3D postures is presented in this research. We use depth maps to infer the positions of 3D skeletal joints. By grouping HOJ3D vectors derived from a vast array of postures, we create posture vocabularies. In this paper, we offer a cutting-edge method for 3D joint location histogram-based human action recognition. On the difficult 3D action dataset, our real-time technique produces improved results.

Cao et al. [11] the difficulty of merging numerous capabilities for motion detection is mentioned in this paper. they devise a unique framework that blends branch-and-bound primarily based detection with STIP illustration primarily based on GMM. The results of the experiments display that this technique can efficiently detect movement even in the presence of a cluttered heritage and partial occlusions. they create a sparkling dataset for the action detection assignment, and their set of rules outperforms the state-of-the-art techniques.

Lao et al. [12] according to the findings of this study, video surveillance can contribute to the protection of humans at home by using easing manage of home-front and gadget utilization functions. We check out a flexible framework for semantic analysis of human behaviour from surveillance video in this paper. Our proposed framework proven excessive quality and close to-real-time overall performance. We proposed a layered framework for multi-degree human movement analysis. The framework captures human motion, classifies its posture, infers semantic activities the usage of interplay modelling, and reconstructs 3-D scenes. The proposed HV-PCA descriptor with temporal modelling achieves an approximate 86% accuracy charge in posture popularity.

III. METHODOLOGY

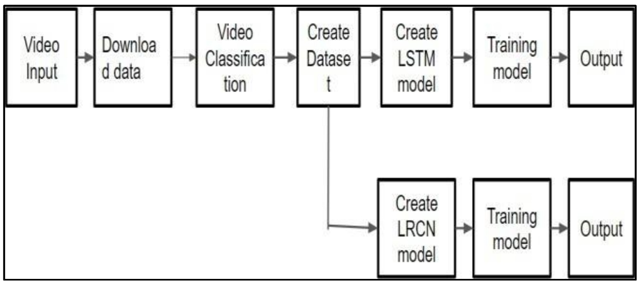

Fig.1. (*Block diagram of HARN system)*

A. Dataset

UCF50 is a hard and fast movement popularity record containing 50 motion classes made from proper YouTube videos. The YouTube pastime dataset, which incorporates 11 movement categories, is improved in these statistics set. the general public of the motion recognition statistics units which are presently reachable are staged by using actors and aren't practical. The UCF50 dataset's main purpose is to provide the laptop imaginative and prescient network get the right of entry to a set of sensible films acquired from YouTube for use in action detection. due to huge versions in digicam motion, item appearance and pose, item scale, viewpoint, cluttered history, and illumination situations, the dataset we used is highly tough. The movies are divided into 25 companies for each of the 50 classes, with every institution containing a minimum of 4 movement clips. The movies inside the same series may all have an equal challenge, a comparable putting, or an identical point of view. Dataset UCF50 50 motion-associated classes were pulled from YouTube, inclusive of Baseball Pitch, Basketball taking pictures, Bench Press, cycling, Billiards Shot, Military Parade, Breaststroke, jumping Jack, easy and Jerk, Diving, Drumming, Fencing, golfing Swing, playing Guitar, high soar, Horse Race, Horse driving, Hula Hoop, Javelin Throw, Juggling Balls, Kayaking, Lunges, mixing Batter, Nun Chucks, Skyjet, soccer Juggling, Swing, desk playing, Taiichi, leap Rope, Tennis Swing, Trampoline jumping, Violin gambling, Volleyball Spiking, dog on foot, and Yo Yo.

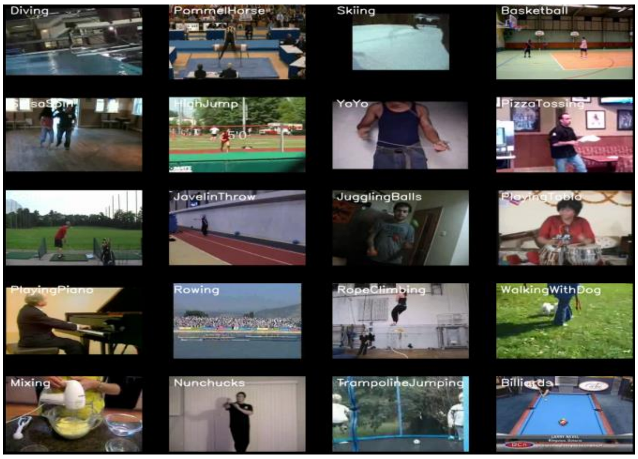

Fig.2. (Data *Creation)*

B. *Convolution Neural Network(CNN)*

A deep neural community known as a convolutional neural community (CNN or Conv internet) is specialized for working with photograph records. It excels at decoding photographs and making predictions about them. As we move deeper into the community, the number of function maps increases and the size of maps has reduced the usage of pooling operations without dropping crucial statistics. it works with kernels (called filters) that test the photo and generate function maps (that represent whether or not a selected feature is a gift at a region in the picture or no longer).

C. *Long Short Term Memory (LSTM)*

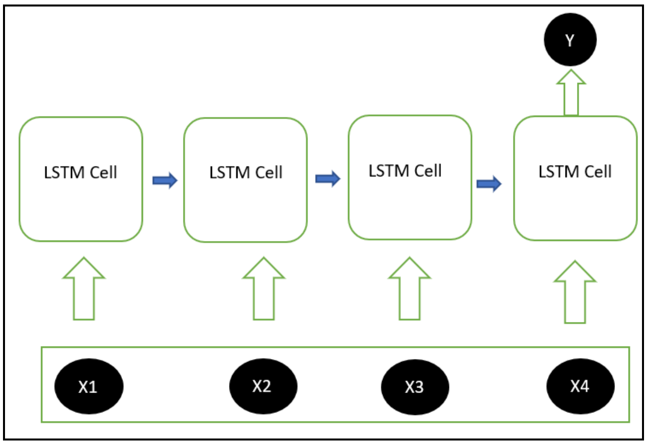

Fig.3. (*Long-short Term Memory)*

In LSTM 3 gates—the enter, output, and overlook gates—are used on this LSTM architecture to tackle vanishing gradient issues and save you lengthy-time period dependency troubles. 3 gates, one for the current time step and the opposite for the output from the preceding step, are used to alter each cell. One time-step at a time, the model will develop the capacity to are expecting human behavior. the reputation of human activity in time collection is a commonplace difficulty. Our sequential method uses a linear stack of two LSTM layers to describe data greater absolutely. so one can save the following LSTM layer from receiving scattered statistics in preference to sequences, the preliminary LSTM layer returns sequences.

D. *CNN+LSTM*

In the enter sequence (video), a CNN will be used to extract spatial features at a selected time step, and an LSTM could be used to decide the temporal relationships between the frames. The ConvLSTM method benefits of the CNN and LSTM architectures are combined in this hybrid model. The counseled layout includes a linear stack of two 1D convolutional layers, two unidirectional LSTMs, and max pooling layers. Convolutional layers will extract the nearby features from the signal records, at the same time as LSTM layers will represent the temporal dependencies on the information. The kernel size for each convolutional layer is 3 x 1 with stride 1. each of the 2 LSTM layers has sixty-four neurons that output sequences. To manipulate computational complexity, dropout is blended with the activation function Relu. The final layer of the CNN+LSTM model, which contains six neurons to recognize six human behaviors, has two linked layers.

E. *LRCN*

Here, we use a different strategy called a protracted-time period recurrent convolutional community (LRCN). in this method, CNN and LSTM layers are blended into one model. so that you can mimic a temporal collection, the convolutional layer extracts spatial statistics from the frame and feeds the extracted spatial information to the LSTM layer at each temporal step. As a result, in the end-to-end mastering manner, the community directly learns spatiotemporal homes.

IV. EXPERIMENTATION AND RESULTS

We used the UCF50 dataset for trying out and education. First, the usage of the benchmark UCF50 we examined the proposed ConvLSTM network to determine its accuracy charge for human movement events. We utilize 25% of the samples as checking out statistics and 75% of the samples as training records for the cautioned community. After education, we tested the data at the LSTM version and used the confusion matrix to calculate the accuracy rate. Then, similarly, for extra accuracy, we tested the information on the LRCN model as well. After comparing the two outcomes, we observed that the LRCN model offers us more accuracy. Then, for greater statistics, we computed its recollect, Precision, and F-score.

The definition of accuracy is the ratio of samples used for proper categorization divided with the aid of the overall number of samples, that's proven as follows:

$$A = TP + TN / TP + TN + FP + FN$$

Bear in mind is described because the actual wonderful rate or sensitivity, which is defined as:

$$R = TP / TP + FN$$

while the equation, which is as follows, can be used to calculate the community's precision;

$$P = TP / TP + FP$$

TABLE I. CLASSIFICATION REPORT OF CONVLSTM

	Precision	**recall**	**f1-score**	**support**
WalkWithDog	0.60	0.60	0.60	30
Taichi	0.91	0.77	0.83	26
Swing	0.70	0.85	0.77	33
HorseRace	0.90	0.84	0.87	43
Basketball	0.72	0.79	0.75	33
Rowing	0.78	0.69	0.73	26
accuracy			0.76	191
Macro avg	0.77	0.76	0.76	191
Weighted avg	0.77	0.76	0.77	191

TABLE II. CLASSIFICATION REPORT OF LRCN

	Precision	**recall**	**f1-score**	**support**
WalkWithDog	0.83	0.80	0.81	30
Taichi	0.96	1.00	0.98	26
Swing	0.86	0.94	0.90	33
HorseRace	0.98	0.95	0.96	43
Basketball	0.97	0.88	0.92	33
Rowing	0.93	0.96	0.94	26
accuracy			0.92	191
Macro avg	0.92	0.92	0.92	191
Weighted avg	0.92	0.92	0.92	191

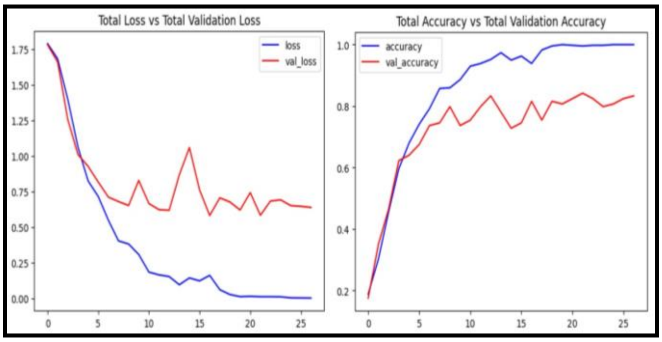

FIG.4. (LSTM *TOTAL LOSS VS TOTAL VALIDATION LOSS & TOTAL ACCURACY VS TOTAL VALIDATION ACCURACY)*

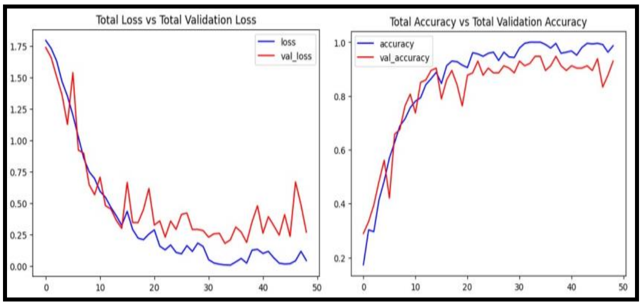

Fig.5. (LRCN Total loss vs total validation loss & Total Accuracy vs Total validation Accuracy*)*

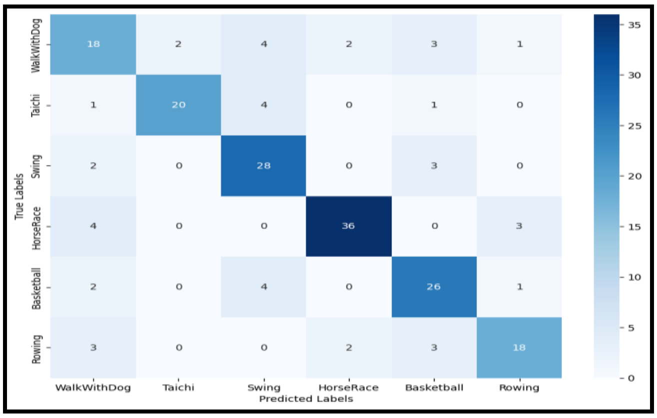

Fig.6. (*Confusion matrix of LSTM Model)*

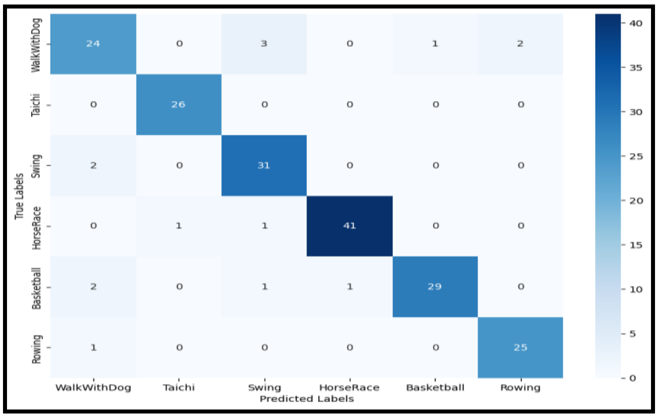

Fig.7. (*Confusion matrix of LRCN Model)*

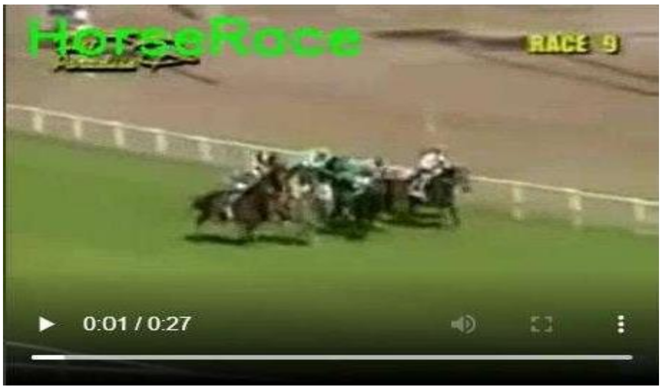

Fig.8. (*Output 1 of HARN system)*

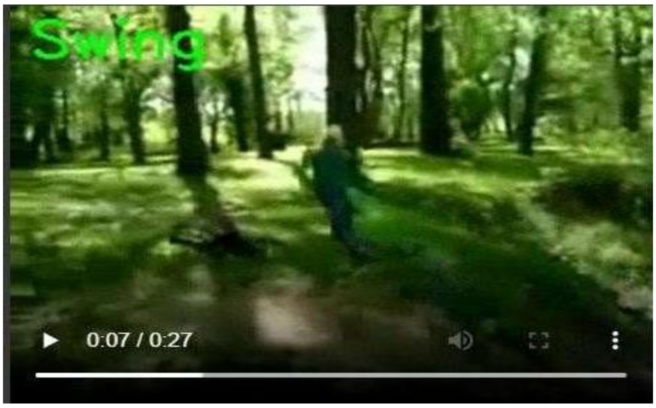

Fig.9. (*Output 2 of HARN system)*

V. CONCLUSION

Considering Computer vision is currently a warm issue, systems like Human hobby recognition systems are rather realistic and efficient for addressing various programs, which include surveillance and tracking, as well as helping the elderly and the blind. further to giving quit customers extra comfort, this can be implemented in a spread of businesses to lighten the body of workers individuals' workloads. The model performs admirably on video streams while doing passably on image streams. modern society places an excessive fee on hobby recognition structures because of the convenience and problems they address. developing demands include the need for activity recognition for monitoring and surveillance, video segmentation, and many others., where this era is probably pretty beneficial. To help the antique and blind even greater, this technique may be covered in mobile apps. even though it saves a tonne of money and time, it's also at risk of human mistakes. This technology serves as the muse for all different activity reputation programs. For widespread or niche goals, this the gadget is therefore very beneficial agency.

ACKNOWLEDGMENT

First and principal, we would like to express our heartfelt gratitude towards the subject faculties. They have been nothing but supportive throughout the course of this project development. Thank you,, for guiding us, answering our questions, clearing our doubts and sometimes even reassuring us. We would also like to extend our thanks to the other faculty members who have been working diligently despite numerous personal and professional difficulties. They have always made us feel that we can approach them anytime asking for help

Lastly, we want to thank our classmates, who showed great enthusiasm and engaged in discussion and debates. Doing our projects alongside each other made this course fun and easy –going.

REFERENCES

[1] Yuanyuan Huang, Haomiao Yang,and Ping Huang, "Action recognition using hog feature in different resolution video sequences," International Co ference on Computer.

[2] Lu, J., Nguyen, M., & Yan, W. Q. (2020, November). Deep learning methods for human behavior recognition. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)* (pp. 1-6). IEEE.

[3] TamV.Nguyen et al.,"Spatial-Temporal Attention- Aware Pooling for Action Recognition," IEEE Transactions On Circuits And Systems For Video Technology, vol. 25, no. 1,pp.77-86, 2015

[4] Yuan, X., & Yang, X. (2009, December). A robust human action recognition system using single camera. In *2009 International Conference on Computational Intelligence and Software Engineering* (pp. 1-4). IEEE.

[5] Reddy, K. K., & Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine vision and applications*, *24*(5), 971-981.

[6] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu ,"3D convolution neural networks for human action recognition", IEEE Transactions On Pattern Analysis And Machine Intelligence , vol. 35, no. 1,pp. 221-231, 2013.

[7] Meng Li, Howard Leung, and Hubert P. H. Shum, "Human action recognition via skeletal and depth based feature fusion, "Proceedings of the 9th International Conference on Motion inGames,pp.123-132,2016.

[8] Dawar, N., & Kehtarnavaz, N. (2018). Action detection and recognition in continuous action streams by deep learning-based sensing fusion. *IEEE Sensors Journal*, *18*(23), 9660-9668.

[9] Zhou, Z., Shi, F., & Wu, W. (2015). Learning spatial and temporal extents of human actions for action detection. *IEEE Transactions on multimedia*, *17*(4), 512-525.

[10] Xia, L., Chen, C. C., & Aggarwal, J. K. (2012, June). View invariant human action recognition using histograms of 3d joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 20-27). IEEE.

[11] Cao, L., Tian, Y., Liu, Z., Yao, B., Zhang, Z., & Huang, T. S. (2010, July). Action detection using multiple spatial-temporal interest point features. In *2010 IEEE International Conference on Multimedia and Expo* (pp. 340-345). IEEE.

[12] Lao, W., Han, J., & De With, P. H. (2009). Automatic video-based human motion analyzer for consumer surveillance system. *IEEE Transactions on Consumer Electronics*, *55*(2), 591-598.

[13] Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document Recognition," Proceedings of the IEEE, pp.2278- 2324,1998

[14] Heng Wang, Alexander Klaser, Cordelia Schmid,and Cheng-Lin Liu,"Dense trajectories and motion boundary descriptors for action recognition,"International journal of computer vision,pp.60-79,2013

[15] Georgios Th. Papadopoulos, Apostolos Axenopoulos and Petros Daras,"Real-time skeleton-tracking-based human action recognition using kinect data," Proceedings of the International Conference ,pp:0302-9743, 2014

[16] Samitha Herath, Mehrtash Harandi,and Fatih Porikli,"Going deeper into action recognition,"A survey of Image Vision Computation, 2017

[17] Jindong Wang et al.,"Deep learning for sensor-based activity recognition,"A Survey Pattern Recognition Letters Elsevier,pp.1-10,2018

[18] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentationand recognition. Computer Vision and Image Understanding", 1152, 2011, pp. 224–241

[19] S.Sadanand and J.Corso,"Action bank: A high-level representation of activity in Video," IEEE Computer Society Conference on Computer Vision and Pattern Recognition,pp.1234-1241. 2012.

[20] K. N. Tran, I. A. Kakadiaris, and S. K. Shah, "Part-based motion descriptor image for human action recognition. PatternRecognition", 457, 2012, pp. 2562–2572