e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 12, December 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712186

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Human Action Recognition in Noisy Videos Using CNNS and Fully Connected LSTMS

Tejaswi Devalla [1], Ananya Donepudi [2], Devanssh Rajhu Cemella [3], Dr. T. Praveen Kumar [4]

Department of AI&DS, Methodist College of Engineering and Technology, Hyderabad, Telangana, India[1,2,3]

Associate Professor, Department of AI&DS, Methodist College of Engineering and Technology, Hyderabad, Telangana, India[4]

ABSTRACT: This study proposes a deep learning model for human action recognition in noisy videos using the UCF-50 dataset. The model combines CNNs for feature extraction and FCLSTMs for capturing temporal information. Preprocessing with batch normalization improves training efficiency. The CNN utilizes max pooling and ReLU activation, and the LSTM has 100 units. A final dense layer with SoftMax activation outputs class probabilities, achieving 90.74% accuracy, demonstrating state-of-the-art performance.

KEYWORDS: HAR, CNN, RNN, Sensors, Spatial data, Temporal data, accuracy, precision, recall, confusion matrix, F1-score.

I. INTRODUCTION

Human Activity Recognition involve using technology like sensors to identify & classify physical movements of a human based on the data collected from various sensors. These sensors can range from simple accelerometers or gyroscopes in smart phones or watches to complex cameras and depth sensors in smart homes and surveillance systems and video cameras for recording movements like walking, sitting, or exercising. Smartwatch sensors for tracking steps, heart rate, and arm swings and mobile phone sensors for measuring acceleration and orientation as you move. The most common datatypes used in this model are:

1. *Video Frames*: A typical approach is to capture the sequence of video frames where each video frame is processed sequentially by the model to capture temporal dynamics in an activity.
2. *Sensor Data*: Different types of data such as 3D acceleration movements or 3D angular velocity can be captured from a sensor.
3. *Combined sensor & video data*: This entails integration of the sensor captured data with extracted features from the video frames to match up with information about movement and context.

Proposed HAR systems faced many limitations that reduce their (i) scalability, (ii) occlusion resulting from background lighting or objects, (iii) class confusion whether from intra-class similarities or inter-class variation.

Among the recent and highly successful approaches to address the challenges and limitations of tackling HAR problems is Fully Connected LSTM. Unlike standard LSTMs that process features at fixed intervals, FC-LSTMs connect all hidden units across time steps. This allows them to learn complex dependencies between distant frames, potentially capturing intricate action details and instigate flexibility.

This combination offers advantages like improved accuracy, in capturing both temporal flow and complex dependencies, FC-LSTMs have the potential to outperform standard LSTMs in HAR tasks. However, a few challenges remain.

(i) *Computational cost*: The full connections within FC-LSTMs can increase computational demands, requiring optimization techniques for practical use. (ii) *Data dependency*: Like other LSTMs, they might still require large amounts of training data, which can be expensive to acquire. Overall, FC-LSTMs present a promising avenue for HAR, but further research is needed to address these challenges and unlock their full potential in real-world applications.

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712186

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. RELATED WORK

[1] **B. Suresh Kumar et al.**: Explores deep learning models (CNNs, RNN-LSTMs) for action recognition on HMDB-51, highlighting SDA optimization challenges and recognition accuracy. [2] **Nenghuan Zhang et al.**: Reviews probabilistic methods, CNNs, and optical flow on AVA and SLAC datasets, emphasizing human action complexity and F1 score evaluation. [3] **Mohanad Babiker et al.**: Proposes a MATLAB-based system using perceptron networks for MSR Daily Activity3D, with static camera limitations and confusion matrix analysis. [4] **Hieu H. Pham et al.**: Discusses CNNs, RNN-LSTMs, and DBNs for UCF-101, focusing on SDA optimization and accuracy improvements. [5] **Djamila R. Beddiar et al.**: Reviews hybrid and deep learning methods for HAR, noting high computational demands and sensitivity metrics. [6] **H. Aman Ullah**: Analyzes DNNs for video-based HAR on Diving-48, addressing privacy concerns and search string limitations. [7] **Jegham et al.**: Implements CNNs, RNNs, and SVMs using TensorFlow on Something-Something V2, focusing on accuracy and recall metrics. [8] **W. Niu et al.**: Explores CNNs, RNN-LSTMs, and HOG features for UCF-101, noting preprocessing challenges and accuracy metrics. [9] **M. Valera et al.**: Compares Bayes Classifier and CNN on HMDB-51, finding CNN superior in accuracy (up to 100%). [10] **N. Albukhary et al.**: Proposes real-time HAR using centroid detection with a single camera, evaluating average processing time and FPS.

III. METHODOLOGY

The proposed architecture focuses on videos by leveraging the strengths of convolutional neural networks combined with fully connected long short-term memory (FCLSTM) networks. The main aim is to efficiently capture spatial and temporal video sequences enabling accurate action recognition. This method integrates a pre-trained CNN for spatial feature extraction and an FCLSTM model for temporal feature extraction. The flow of the proposed architecture is depicted in *Fig-1*. The CNN processes each frame to produce high-dimensional feature maps, which serve as the input for the subsequent FCLSTM network. We used ReLU (Rectified Linear Unit) activation function in CNN to introduce non-linearity, allowing the network to learn and model complex patterns and features in the data. ReLU also mitigates the vanishing gradient problem, enabling faster and more effective training by promoting sparsity in the activation.

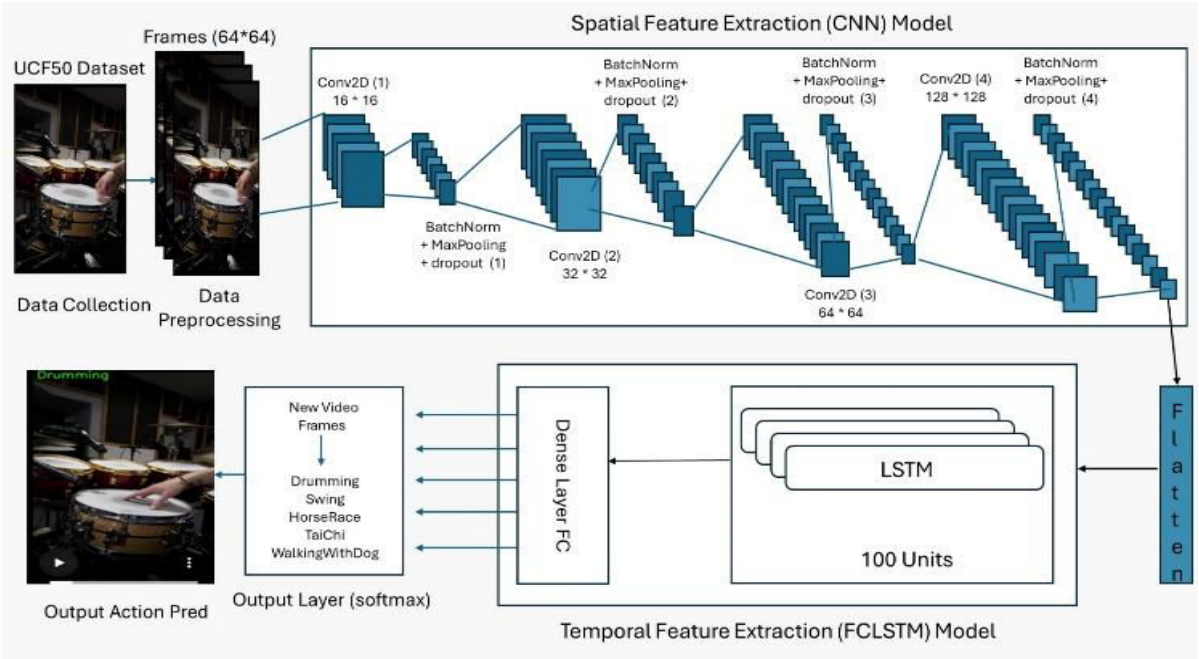

***Fig-1**: Proposed Architecture of CNN+FCLSTM for Human Action Recongnition in videos*

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

FCLSTMs are chosen for their proficiency in handling sequential data, making them well-suited for modelling the temporal dependencies between video frames. The FCLSTM network comprises several LSTM layers followed by fully connected layers, allowing it to learn the sequence of actions over a long time. This combination enables the architecture to understand and predict complex action sequences, offering a robust solution for human action recognition in videos. The LSTMs are a combination of four gates namely, forget gate, input gate, cell state, output gate as explained below.

- *Forget Gate:* Determines what information from the cell state should be discarded.

$$f_t = \sigma\left(w_f \cdot [h_{t-1}, x_t] + b_f\right)$$

- *Input Gate:* Controls the extent to which new information is added to the cell state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

- *Cell State:* Maintains long-term memory of the network, updated by both the forget and input gates.

Update: $\underline{C_t} = tanh\ tanh\ (W_c \cdot [h_{t-1}, x_t] + b_c)$; New Cell State: $C_t = f_t * C_{t-1} + i_t * \underline{C_t}$

- *Output Gate:* Determines the output of the LSTM unit and how much of the cell state should be revealed to the next layer.

Output Gate: $o_t = \sigma(W_0 \cdot [h_{t-1}, x_t] + b_0)$; Hidden State: $h_t = o_t * tanh\ tanh\ (C_t)$

Initially the CNN is employed to process individual frames from a video sequence. The CNN's primary role is to extract detailed spatial features from each frame capturing patterns such as edges, textures and complex shapes. This significantly enhances the feature extraction process and reduces the amount of training time required.

Convolutional Layer (Conv2D):

$$Conv2D = ReLU(Conv(x, W) + b)$$

Rectified Linear Unit (ReLU):

$$ReLU(x) = max\ (0, x)$$

Here, we used batch normalization to standardize the learning process and improving the training speed and

performance. The formula is as follows:

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}, y = \gamma\hat{x} + \beta$$

For feature extraction, we used max pooling which reduces the dimensionality of feature maps by taking the maximum value over a defined window, retaining the most significant features and reducing computational load. The formula is as follows:

$$y = \left(x_{i,j}\right)$$

Once CNN processes each frame, it outputs high dimensional feature maps, containing the spatial features of input frames providing a detailed representation of the visual information, Softmax activation is used in the last layer if CNN for multi-class classification problems as it provides a probabilistic interpretation of the model's output, making it easier to determine the predicted class by selecting the highest probability. These maps are now fed to FCLSTM network.

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\Sigma j e^{x_j}}$$

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712186

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Sigmoid activation is used in the output layer of CNNs for binary classification problems, providing a probability that the input belongs to the positive class, and it can also be used in hidden layers to introduce non-linearity.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Thus, CNN and FCLSTM together provide an efficient model that is excellent at both temporal and spatial analysis. This dual capability is particularly advantageous in real-world applications where actions need to be recognized reliably across diverse and dynamic scenarios.

IV. EXPERIMENTAL EVALUATION AND ANALYSIS

A) DATASET: UCF-50 DATASET

The UCF-50 dataset features 6,618 video clips from YouTube, categorized into 50 human actions (walking, guitar playing, etc.). With at least 100 videos per category, it offers a rich resource for training and evaluating action recognition models. Videos exhibit variations in camera angles, lighting, and backgrounds, mimicking real-world complexities. Labeled with actions, UCF-50 facilitates supervised learning tasks.

Fig-2: *UCF-50 Human Action Recognition Dataset*

B) EVALUATION METRICES

i) ACCURACY

Accuracy of an algorithm is represented as the ratio of correctly classified predictions (TP+TN) to the total number of predictions (TP+TN+FP+FN).

Accuracy = (TP+TN) / (TP+TN+FP+FN)

ii) PRECISION

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as the ratio of true positives to the sum of true and false positives.

Precision= TP / (TP+FP)

iii) RECALL

Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

Recall = TP / (TP+FN)

iv) F1 SCORE

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

Recall = TP / (TP+FN)

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712186

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. OUTPUT

A) EVALUATION METRICES

The evaluation metrics reveal promising performance for the model on most action classes. "Drumming" achieved a perfect score (precision, recall, and F1-score of 1.00), indicating the model flawlessly identified all drumming videos. "TaiChi," "HorseRace," and "Swing" also exhibited strong performance with F1-scores exceeding 0.90, suggesting a good balance between precision and recall for these classes.

Classification Report:

	precision	recall	f1-score	support
WalkingWithDog	0.85	0.69	0.76	32
TaiChi	0.90	0.93	0.91	28
HorseRace	0.90	0.93	0.92	29
Swing	0.88	0.97	0.92	37
Drumming	1.00	1.00	1.00	36
accuracy			0.91	162
macro avg	0.90	0.90	0.90	162
weighted avg	0.91	0.91	0.90	162

***Fig-3**: Evaluation Metrices for 5 classes*

"WalkingWithDog" showed a slightly lower F1-score (0.76) due to a higher recall (0.69) compared to precision (0.85). This implies the model might be capturing most of the walking-with-dog videos but could be misclassifying some videos from other classes as "WalkingWithDog". Overall, the high accuracy (0.91) and macro/weighted averages support the model's effectiveness in recognizing various human actions.

B) CONFUSION MATRIX

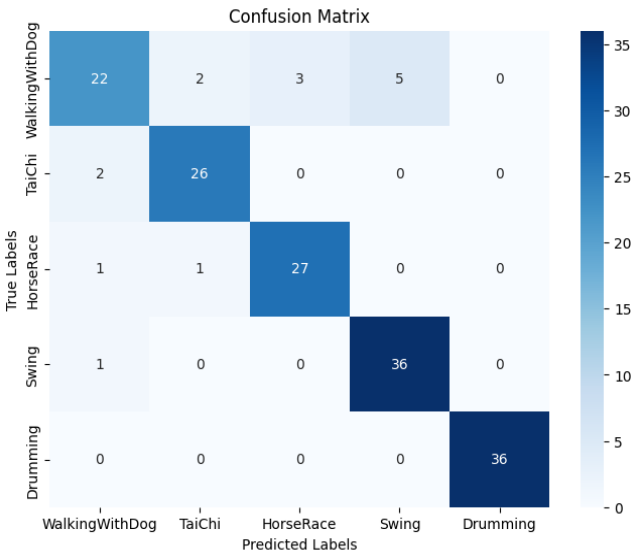

***Fig-4:** Confusion Matrix of CNN + FCLSTM (5 Classes)*

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712186

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C) OUTPUT

***Fig-5**: Output of video action Prediction on frames*

The image depicts the output of our CNN+FCLSTM model successfully recognizing the action in a noisy video. Despite the noisy and blurred quality of the video, the model accurately identifies the activity. Below the frame, there's a confirmation of the predicted class along with the processing time, of 65 milliseconds per step. This demonstrates the model's capability to process and classify the video frames efficiently and correctly, even under challenging visual conditions.

The image (Fig-6) shows the output of our CNN+FCLSTM model correctly recognizing an action within a single frame of a video. At the top of the image, the label "Drumming" is overlaid in green text, indicating the predicted class of the action. This shows that the model has successfully identified the activity occurring in the frame.

***Fig-6**: Output prediction on Single Frame*

No. of classes	With Batch-Normalisation	Without Batch-Normalisation
4	94.35%	81.14%
5	90.74%	--
8	79.59%	--

***Table1**: Performance comparison with and without batch normalization having kernel size (3,3).*

No. of classes	Kernel (3,3)	Kernel (4,4)
4	94.35%	82.73%
5	90.74%	79.35%
8	87.04%	--

***Table2**: Performance comparison with batch normalization in different kernel sizes.*

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D) MODEL'S ACCURACY & LOSS CURVES

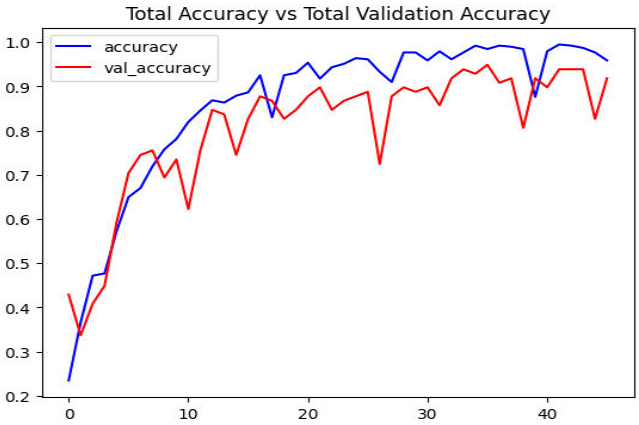

Fig-7: *Accuracy graph of CNN + FCLSTM Model for 5 Classes*

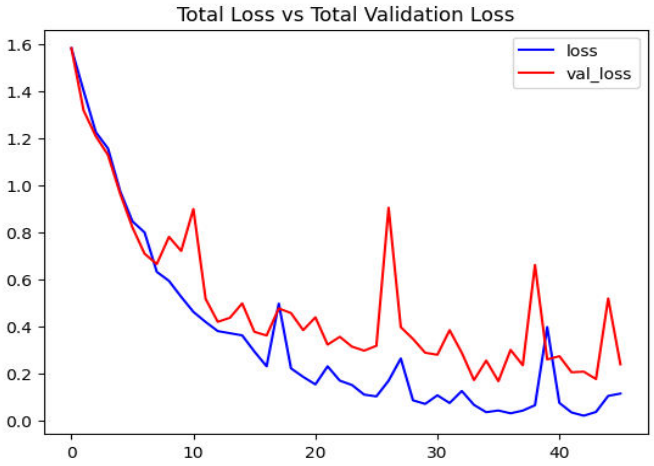

Fig-8: *Loss graph of CNN + FCLSTM Model for 5 Classes*

The graph showcases promising performance for a CNN-FCLSTM model on 5-class action recognition. Training accuracy rapidly rises to near perfection, while validation accuracy climbs to 0.9 with minor fluctuations. The small gap between curves suggests good generalization and minimal overfitting. This indicates the model effectively learns from the data and performs well on unseen examples.

The CNN-FCLSTM model's loss curves (training & validation) depict a downward trend, indicating successful training and potential for generalization. While validation loss fluctuates slightly, the overall decrease suggests the model is learning without overfitting. A wider gap between the curves, however, would warrant concern about generalization ability.

E) CONCLUSION AND FUTURE WORK

Human Activity Recognition (HAR) has witnessed an exciting journey with machine learning, advancing from basic feature extraction to sophisticated deep learning models. However, challenges remain. Overfitting and limited diversity in data hinder generalization to unseen scenarios. Recognizing actions in complex environments with occlusions and

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

varying lighting conditions have been difficult. Additionally, large training data requirements create concerns about scalability of the modal.

In this proposed work, we developed and tested CNN + FCLSTM models for action recognition on the UCF50 dataset, focusing on classes "WalkingWithDog", "TaiChi", "HorseRace,", "Swing," and "Drumming" each containing more than 100 videos. The models were trained using a batch size of 4 over 50 epochs with the Adam optimizer and categorical cross-entropy loss function. The incorporation of batch normalization significantly improved model performance, as evidenced by the highest accuracy of 94.35% achieved by the CNN + FCLSTM model trained on four classes with batch normalization. However, the model showed lower performance without batch normalization and faced challenges when scaling up to more classes. Comprehensive testing, including unit, integration, functional, system, and acceptance testing, confirmed the model's reliability and highlighted areas for optimization, such as handling high collision rates and excessive noise. While the model performed well under normal conditions with just fine noisy data, it struggled with unclear noisy data.

Additionally, GPU disconnections occurred when processing more than eight classes, indicating the need for further refinement to improve computational efficiency and robustness. Future work will focus on these improvements to enhance the model's scalability and performance under challenging conditions. Thus, continued optimization and testing are essential to ensure the model's practical applicability and reliability in diverse real-world scenarios.

REFERENCES

[1.] Deep learning approaches for human action recognition authors: Badhagouni Suresh Kumar1, S. Viswanadha Raju2 and H.Venkateswara Reddy
[2.] Real time human activity recognition with video classification (2022)
authors: S Jahnvi, Chandra Shekar Malepati
[3.] A Review of human activity recognition in videos authors: Nenghuan Zhang,Yongbin wang,Peng Yu
[4.] Intelligent Video Surveillance System (2019) authors: Mohanad Babiker, Othman O Khalifa, Kyaw Kyaw Htike, Aisha Hassan,Muhammed Zahradeen
[5.] Deep Learning Approaches for Human Action Recognition (2022) authors: Hieu H Pham, Alain Crouzil,Pablo Zegers
[6.] State of the art human activity Recognition (2020) authors: Djamila Romaissa Bediar, Brahim Nini, Mohammad Sabroku, Abdenour Hadid
[7.] Analysis of Deep Neural Networks for human activity recognition in videos (2021) authors: H Aman Ullah
[8.] Vision based human activity Recognition on pre trained Alex net (2019) authors: Norul Najirah Mohammed Zamri, Pang Ying Han, Goh Fan Ling, Ooi shih Yin
[9.] Human activity recognition using bayes classifier and convolutional neural network (2018) authors: Congcong Liu, Jie Ying, Feilong Han, Ming Ruan
[10.] Generating videos with scene dynamics - NIPS (2016) authors: Hieu H Pham, Alain Crouzil, Pablo W Niu J Long, D Han YF Wang M Valera SA Velastin
[11.] Real time human activity recognition Using Visual System (2017) authors: N Albukhary, YM Mustafah
[12.] Human Activity Recognition using tools of Convolutional neural networks (2022) authors: MD Milon Islam, Sheikh Nooruddin Fakhri Karray Ghulam Mohammed
[13.] Advancements in Human Activity Recognition: AI models and Applications (2022) authors: Neha Gupta, Suneet K Gupta.
[14.] Human Activity Recognition using Attention - mechanism based deep learning feature combination (2023) authors: Morsheda Akter, Shafew Ansari, MD Al Masrur khan, Dongwan kin
[15.] Human activity Recognition using convolutional LSTM with spatial temporal networks (2019) authors: Ashok sarabhu, Ajit kumar santra
[16.] Human action recognition using two stream attention-based LSTM networks (2019) Authors: Cheng Dai, Xingang Leu, Jinfeng Lai
[17.] Learning Long term temporal features with deep neural networks for human action recognition (2019) authors: Sheng Yu, Li Xie, Lin Liu, Daoxun Xia

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[18.] Acknowledgement of patient in sense behaviours using bi directional Conv LSTM (2023) Authors: Upendra Singh, Puja Gupta, Mukul Shukla, Varsha Sharma, Sunita Varma, Sumit Kumar Sharma
[19.] Human activity recognition using convolutional neural networks and kinetics dataset (2020) authors: Ms. Shikha, Rohan Kumar, Shivam Aggarwal, Shrey Jain
[20.] Human activity recognition and motion analysis (2011) author: JK Aggarwal
[21.] Investigation on human activity recognition using deep learning (2022) authors: Velliangiri Sarveshwarana, Iwin Thankumar Joseph, Maravarman Mc , Karthikeyan Pd
[22.] Enhancing video recognition with spatial temporal pyramid networks (2019) authors: Zhenxing Zheng, Gaoyun An, Dapeng Wu
[23.] Human activity recognition via hybrid deep learning (2021) authors: Imran Ullah Khan, Sitara Afzal, Jong Weon Lee
[24.] Attention Based Convolutional LSTM for describing video (2020) authors: Zhongyu Liu, Tian Chen, Enje Ding, Yafeng Liu, Wanli Yu
[25.] A novel human activity recognition architecture using residual inception Conv LSTM layer (2022) authors: Sarah Khater, Mayada Hadoud , Magda B Fayek

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com