International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

Human Activity Recognition using LSTM with depth data

[1]Kumari Priyanka Sinha, [2]Prabhat Kumar, [3]Rajib Ghosh

Submitted: 26/05/2023 **Revised**: 06/07/2023 **Accepted**: 23/07/2023

Abstract: For many academics, HAR is a hot topic. They can do this with ease because to a number of cutting-edge technologies, including deep learning, which is useful in a number of contexts. While most of the current body of work has focused on wearable sensor data, it is not always practical to get such data. Publicly accessible video datasets are mined for human activity detection in the proposed study using deep learning techniques including CNNand long short-term memory. CNN extracts relevant characteristics from input data, whereas LSTM eliminates and rejects superfluous data to increase performance. The confusion matrix's precision and recall are used to evaluate the suggested technique. Accuracy is high across the board, as shown by the fact that the diagonals of the confusion matrices for all actions are near to 1.

Keywords: *HAR, deep learning, CNN, neural network*

1. Introduction

The ability to identify human actions is now fundamental to survival. Classifying a person's actions in real-time from a series of sensor data [1]or visual data collected from a variety of input sources is the challenge known as human activity recognition. To accomplish its goal of identifying human behaviour, activity, or condition, human identification systems use data from a wide variety of sources[2].As wireless data transmission, Bluetooth, and cellular data continue to advance, this data may be quickly moved to a new media and put to use in modelling. The human's motion may be tracked in real-time[3] without any lag. In spite of its rising popularity over the last decade, there are still many obstacles to overcome before it can reliably and quickly translate raw input data into well-defined, actionable motion. The identification of human actions from still or moving photographs is a difficult problem. Problems with size, clutter, occlusion, perspective, lighting, and overall presentation must be addressed.

Video surveillance, human-computer interaction, and human behaviour recognition are just a few examples of the many uses for multi-activity recognition.

Developing an automated system that can properly identify the activity being done by a person by evaluating data from a variety of sources is the primary goal of HAR. There may be variations in the procedure based on the nature of the data source, the nature of the input data, the model training architecture, the kind of activities being performed, and the intended use of the system.

It has proven useful in a variety of settings, from automated surveillance to healthcare to elder care to sports to robots to security to media broadcasting, and it has all showed great promise in solving real-world issues by integrating technology. In addition [4], the use of vision datasets for the sake of preventing potentially risky actions and identifying criminals is a major application.

The use of human activity has demonstrated promising results and substantial needs in the field of automated video surveillance [5]. “These systems are ideal for intelligent crowd surveillance in shopping malls, games, live concerts, streets, and highways, crossroads, traffic lights, parking lots, since they can easily identify unwanted and suspicious activities and track people in the crowd.” To recognise moving objects, a convolutional neural network with an LSTM model is fed frames from a video feed as input [6] before being trained to extract temporal and spatial features. As an added bonus, this may assist mitigate risks by decreasing response times, balancing the burden of security staff, and immediately alerting the appropriate parties. Redundant frame detection [7] applied as pre-processing using a convolutional neural network may be used for successful results in classifying various human activities. Classifying human actions is aided greatly by the pre-processing of video frames.

Image/video from cameras, video recording devices, surveillance cameras, 3D cameras, Microsoft Kinect cameras, infrared sensors, etc. is used for activity detection in vision-based activity recognition[8]. A kinetic camera can take depth shots, however a standard

[1]*Department of Computer Science and Engineering,*
Nalanda College Of Engineering, Chandi, India
kumaripriyankas.phd18.cs@nitp.ac.in
[2]*Department of Computer Science and Engineering,*
National Institute of Technology Patna, Patna-800005, India
[3]*Department of Computer Science and Engineering,*
National Institute of Technology Patna, Patna-800005, India

camera can only take 2D or 3D photos. There's a lot of fresh visual data thanks to security cameras and YouTube, but there are also issues. These include, but are not limited to: background clutter; partial occlusion; viewpoint bias; inconsistent lighting; awkward camera angles; and shadows.

Despite the importance of video surveillance, there are several obstacles [9] to overcome when trying to identify human behaviours inside recordings, such as rapid view shifts, a lack of adequate view angles, etc. Information loss over a longer period of time might be the outcome of these difficulties. Even if it is difficult to link the target over spatial proximity in the case of camera position change. Associating objectives in an interval with little available information calls for a fresh strategy[10].Untrimmed depth recordings make it difficult to identify and classify posture changes. When it comes to the classification of posture change segments or body movements, [11] uses CNN to extract characteristics from video frames.

The majority of studies using human activity recognition have used eyesight or wearable sensors. Problems arise when trying to use these wearable sensors [12] for HAR, since they need to be connected to the subject's body, which isn't always possible. There is an alternative to employing wearable sensors for HAR: using video frames captured by cameras.

2. Deep Learning

Artificial neural networks (ANNs) replicate human brain function. Neural networks learn from training input to predict or classify output. Input and output data are analysed by the ANN architecture, which then finds patterns and correlations. Input layers, hidden layers, and an output layer make up an ANN. For image identification tasks, neural networks excel. For this purpose, CNNs stand out as the optimal neural network architecture. Deep learning is a subfield of artificial neural networks that has found use in a variety of problem-solving contexts, including but not limited to the ones listed above. In deep learning [13], a unified algorithm with a single kind of activation function is used to handle input at each successive layer. Data characteristics useful for instruction, discovery, and comprehension are built layer by layer.

3. Discriminative Deep Learning Models

Discriminative feature learning models are demonstrated with consequent distributionclasses to boost their classification and recognition powers. CNNs, RNNs and other discriminative deep learning are used to identify human behaviour.

3.1 Convolutional Neural Network （CNNs）

Deep learning CNNs can detect and extract attributes from an input image using learnable biases and weights. CNN[14]'s ability to collect both temporal and spatial connections in an image and minimise the picture without sacrificing aspects that help build a more scalable prediction model are its main strengths.A conventional CNN has two primary parts. As illustrated in Figure 1, there are two main components: feature extraction and classification.

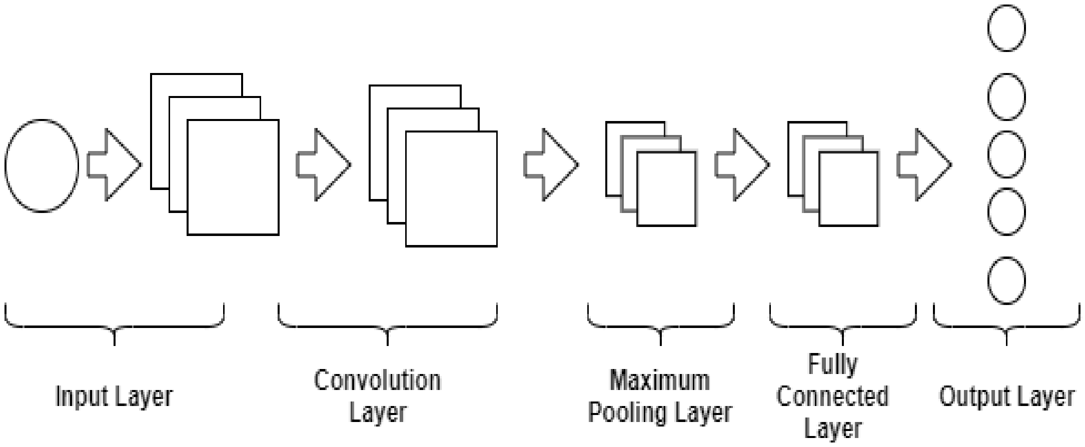

Fig 1: CNN Architecture

Standard feature extraction architecture [2] contains input, convolution, or pooling layers. Convolution layer extracts data from image or video frames [15]. A filter in the layer of convolution is gradually dragged across the input image to convolve it. The feature map or activation map is the result of the convolution layer.

- The 1D or 3D convolution stage is selected based on the kind of input image[16]. Basic picture characteristics like edge and corner colour, gradient orientations, and so on are extracted in the first layer. To further extract features from the input picture, this will be fed into higher-level layers. High-level details about the input picture are extracted by the deeper convolution layers. The equation for the convolution layers goes like this.-

$$x_i^{l,j} = f\left(\sum_{a=1}^{m} w_a^j \;\; x_{i+a-1}^{l-1,j} + b_j\right)$$

- One-dimensional or three-dimensional convolution, depending on the input picture type[16]. It is the job of the first layer to extract the most basic aspects of the picture, such as the colours and orientations of edges and corners and any gradients present. This will be sent into further layers to help extract more information from the input picture. When combined with the information from lower-level convolution layers, the high-level features extracted by the higher-level convolution layers reveal the whole context of the input picture. Formulation of the convolution layers.
- • Fully connected layers are used to predict the output labels or classes of an input image by flattening the features recovered during feature extraction. The neurons in this layer are taught to perform non-linear functions by adjusting the weights and biases between them. It's a useful tool for maximising outcomes like test scores. The next layer, the output layer, will use SoftMax classification algorithms to assign labels to the images. The dropout layer is another popular layer in CNN architecture. By eliminating part of the network's neurons during training, this layer may assist alleviate the overfitting issue.

Remarkable progress in artificial intelligence has been made thanks to the development of many CNN architectures during the last few of decades.

3.2 Recurrent Neural Network

The usage of RNNs for sequential input in natural language processing has increased in recent years. A recurrent neural network utilises its previous output as input and contains hidden states. “Voice recognition, text recognition, speech recognition, and forecasting require time-series sequential data. [17] An RNN can calculate the current state using the current input and the previous state's output and uncover the relationship between current and prior inputs.” Thus, it contains at least one feedback link, allowing activation to circulate in a closed loop. As a result of the hidden state, RNNs are very well-suited to processing time-domain data, making them ideal candidates for the HAR dataset.

At each time step, the RNN model's gradients disappear and extend to zero and infinity [17]. In gradient back-propagation, the weight matrix multiplies the gradient signal many times. The gradient value will be driven to zero by repeated multiplication if the weight matrix's eigen value is less than one, causing the vanishing gradient problem. When the weight matrix is larger than one, the value is pushed to infinity, leading to exploding gradient issues. LSTMmodel solves this issue by introducing a novel component called a memory cell.[18].

3.3 Hybrid Network

When two or more classifiers are combined, the resulting network is called a hybrid network [19]. It combines the predictions of two or more models, trains them together, and integrates the results for improved accuracy. In the realm of Machine Learning, this strategy is referred to as "ensemble learning." When attempting to anticipate the same issues using numerous models, this method takes an average of the results. Combining CNN and LSTM[18] models lets you learn spatial and temporal data simultaneously. “A CNN creates high-level spatial information on the activity of the photos for image classification, while an RNN model extracts temporal correlation between the clips' frames by remembering the preceding frames. In addition to ensembles of numerous models, hybrid approaches combine form-based and motion-based properties to depict an action”. Optical flow or a histogram of motion intensity are used to capture motion data, while shape-based features are extracted from the still picture for use in action detection.

4. Working Model and Proposed Methodology

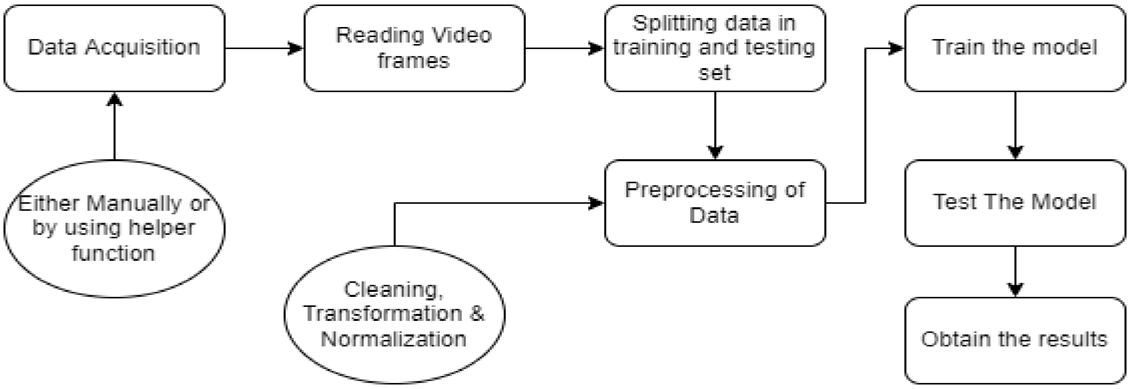
Data Acquisition
Reading Video frames
Splitting data in training and testing set
Train the model
Either Manually or by using helper function
Preprocessing of Data
Test The Model
Cleaning, Transformation & Normalization
Obtain the results

Fig 2: Working Model for Human Action Recognition

Convolutional and recurrent neural networks are commonly used in human action recognition models. To minimise spatial dimensions and boost input array depth, the working model will combine Convolutional and Maxpooling layers[20]. We will next add a series of fully connected layers on top of this, with the last (output) layer consisting of six neurons, one for each of the enumerated operations. A likelihood that the input video belongs to a certain category will be provided by the model. The most likely classification label for that video is the one we'll use to describe it. There will be two distinct sets of information, one for training and one for testing. The model will be educated with familiar information from the training set, and then put to the test with information it has never seen before. After collecting data, the video model extracts just the frames it needs in a predetermined manner, which is then fed into a machine learning algorithm so that it can make a more accurate forecast.

Recent years have seen the rise of deep learning as a popular and very efficient method for human activity identification in video. Convolutional neural networks (CNNs) are employed in the proposed study as a deep learning technique for extracting features from video clips. When training a model for classification, the convolutional layer may be used to extract and learn the features that are then utilised to classify data. The suggested convolutional neural network (CNN) model can learn spatial characteristics from a single video frame. The geographical information is captured well by CNN, but the temporal data is not. For human activity identification, temporal data from video sequences is also crucial for capturing motion.

Convolutional layers let CNN extract characteristics from input video frames while maintaining key information. CNN is utilised for feature extraction, and the weights are determined by the pre-training processing of its specialised neural network networks. The input picture is utilised by the feature extraction network during extracting features. The neural network performs classification based on the retrieved characteristics.

Using the input data and the CNN filter, the convolutional layers generate the output features.

The input x (i, j) is convolved with the filter w (i, j) in the convolutional layer. C d is the filter size, which is recorded in z. (i, j)-

$$z(i,j) = x(i,j) \times w(i,j) = \sum_{a=-c}^{c} \sum_{b=-d}^{d} x(a,b).w(i-a,j-b)$$

In the process of action recognition using a convolutional neural network (CNN), features are retrieved that include spatiotemporal information. Recurrent neural networks, including LSTM in particular for pre-processing of video sequences to learn and analyse the temporal aspects for human activity detection, are used to manage the spatiotemporal data. The LSTM network is used to eliminate extraneous details from the input data. However, the suggested deep learning method works well for temporal information over relatively short periods of time but is not suited for longer sequences. In reality, the LSTM network is used to categorise sequential input, which may then be utilised for action recognition.

Bidirectional LSTM networks classify using extracted characteristics. Certain instances of video classification for action recognition employ transfer learning [21]. Transfer learning applies the insights learned from training a model on a large dataset similar to a new dataset. As a result, the characteristics learnt in the training set are being put to the test in the validation set. Creating a network capable of handling massive amounts of video data requires a tremendous amount of storage space and processing power.

5. Implementation

5.1 Dataset

When it comes to HAR studies, datasets are crucial. The primary goal of HAR is to determine what a person does based on information gathered from many sources. Numerous research have been conducted to better organise and connect this mountain of data with relevant facts and insights. A dataset that details behaviour in a variety of contexts is necessary for the study's success. Therefore, the dataset's availability and quality are crucial to training a model to accurately recognise activities.

Numerous researchers have made use of the various publicly available datasets that have been created and released. Using the same datasets allows for a more accurate comparison of various methods and an assessment of their effectiveness. Unlike real-world situations, many datasets were captured in standardised experimental settings with fixed cameras and backdrops.

Datasets can be broken down further into subcategories based on a variety of criteria, including the type of data collection method used [22] and the type of activity being studied. KTH's Activity Dataset was utilised for this study. This dataset, which has been around since 2004, has 25 people engaging in 6 distinct activities across 4 distinct controlled environments. A stationary camera films one person walking, jogging, running, boxing, waving, and applauding. Its views, stories, and action-packed activities vary.

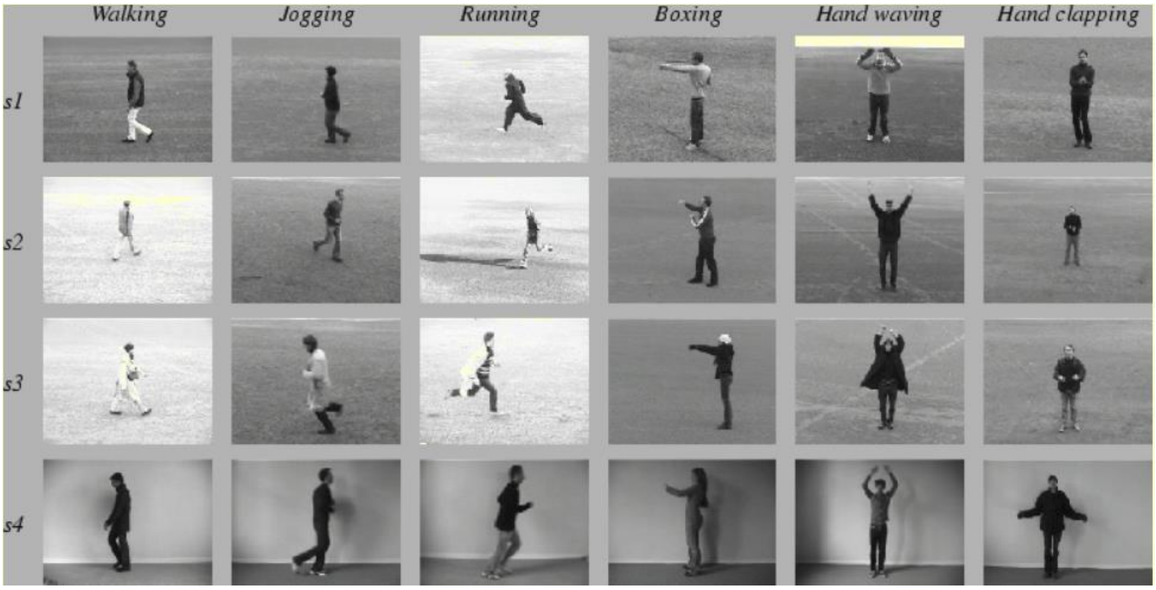

Fig 3: KTH [16] Activity Dataset

5.2 Data preparation

TensorFlow with the Keras backend was used to create both the CNN and the RNN. RNN's input format is distinct from that of CNN's. Each RNN data sample has to be in the form of a data sequence. In the context of the study at hand, a single data sequence corresponds to a single trial of an activity done by a single volunteer. RNN requires uniformly sized input data samples. However, the sizes of the individual data files vary. Cropping is done before RNN input data files are of the same length.

5.3 Results and Analysis

We conduct experiments on the KTH activity dataset and analyse the results using a model that has been pre-trained using the KTH dataset. The remaining 30% is utilised for testing after 70% of the data has been used for training. Python's OpenCV package is used to extract dense optical flow. The model's deep learning component is implemented using Kears. For every action in the KTH activity dataset, confusion matrix evaluates model performance.

To attain the required performance accuracy in machine learning, it is crucial to develop an effective machine learning model assessment metrics. Each kind of challenge requires a unique set of criteria for examination. Different metrics are used for classification, regression, ranking, clustering, associating, etc. “The evaluation metrics not only provide the parameter by which the model's performance can be gauged, but also aid in explaining the results obtained with alternative implementations. There are a variety of criteria used to evaluate the efficacy of machine learning models.”AOC Many accuracy measurements may be determined, including the F1 score, mean absolute error, and mean squared error.

The enigma matrix looks like-

Table 1: Matrix of Confused

Actual Class		Negative	Positive
	Negative	TN	FN
	Positive	FP	TP

TP are cases that have been accurately identified as positive, whereas FP are those that have been misidentified as negative. Truly negative examples, or "True negatives," are those that have been appropriately labelled as such. FNare the opposite of true negatives (DN).

Each action in the dataset is given a precision and recall based on the aforementioned values.

Table 2: Precision and Recall for each activity

Activity	Precision	Recall
Walking	0.93	0.95
Jogging	0.80	0.82
Running	0.90	0.85
Boxing	0.94	0.90
Hand waving	0.97	0.94
Hand Clapping	0.84	0.91

Table 3: Confusion Matrix

True Class

Walking	0.95	0.60	0.03	0.00	0.00	0.00
Jogging	0.50	0.82	0.00	0.00	0.00	0.00
Running	0.15	0.40	0.85	0.00	0.00	0.00
Boxing	0.00	0.10	0.31	0.90	0.00	0.00
Hand waving	0.10	0.00	0.00	0.00	0.94	0.30
Hand Clapping	0.00	0.00	0.00	0.00	0.00	0.91

Predicted Class

The suggested model's confusion matrix for all six actions is shown in Table 3. All of the diagonal entries in the aforementioned confusion matrix have values extremely near to 1, suggesting that the accuracy of the predictions for each action is high.

6. Limitations

Recognising human actions has been a hot topic in the scientific community recently. The widespread application of HAR to industries as varied as sports, video surveillance, filmmaking, and medicine is also a contributing factor. In order to recognise actions, several techniques have been tried. The effectiveness of sensor and video data for HAR depends on devices, data quality, experimental settings, light fluctuations, moving background, viewpoint shift, occlusion, noise, and so on. There is a discussion of some of the difficulties and restrictions of HAR here.

- Video datasets are notorious for their high memory requirements. As a result, it would be impracticable to put the complete dataset into RAM. Some solutions have been presented to this problem, such as passing a URL to download a video from a website like YouTube. Another issue is that there isn't a universal benchmark dataset that covers all the bases in terms of representing real-world circumstances and behaviours. Because of this, HAR assessment and training have become less reliable. More work is required to compile a reliable collection of accurate data, and a standard approach should be used to conduct quantitative comparisons across different benchmarks.
- • The intra- and inter-class differences in HAR provide a significant difficulty. Subjects' experiences of the same task might vary depending on their size, attire, and character traits. The way a person walks, for instance, may be completely unique to that individual. In addition, certain physical activities, such as walking or jogging, may appear remarkably similar to others. It may be challenging to teach a system to recognise complex behaviours that include numerous activities, such as sipping tea while conversing on the phone. More precise data on these actions and activities enables a deep learning model to discern between them, and hybrid devices can train a multileveled prediction model to identify composite activities.

- Dynamic backgrounds, occlusions, illumination variation, noise, varying lighting, varying perspectives, and low-quality photos and videos are all commonplace in real-world videos. Noise, undesired signal detection, and subpar sensors and gearbox systems may all compromise the quality of data acquired by sensor-based methods. These factors increase the difficulty and difficulty level of HAR. Sensor-based HAR may make use of multi-sensor data, as well as data from other sources, such as RGB, depth, a video skeleton, and more.
- A HAR system may use more energy and materials. Real-time precise sensing is essential for many uses, including video monitoring and care for the elderly. More processing power, electricity, and memory will be required to process the massive amounts of data generated by sensors and videos. Some programmes, including those concerned with security, need to be able to foresee a user's next move based on an analysis of their past actions. Because of these constraints, automated activity identification in real time is now more difficult than ever.

7. Conclusion

Computer vision, robotics, and various other applications use Human Action Recognition (HAR) to analyse and interpret human activities. Accelerometers, gyroscopes, and magnetometers in smartphones, smartwatches, or video security cameras make data collection easy. Our study makes use of a publicly accessible activity dataset including a variety of body parts and positions.

We examine the information and sort people into categories using methods like machine learning and deep learning architecture. Researchers showed that raw data alone might provide better results if a balanced dataset is utilised for training the model. Activity detection and categorization using video is useful in certain domains, such as sports. We employed a state-of-the-art method that relied on a sports-themed model that had already been trained to do this. Our dataset may be expanded to include the classification of scoring actions linked to individual sports, in addition to the classification of various sports based on activity conducted by the person in the video. The findings have established a baseline performance for output, and the hard dataset will aid researchers in classifying computer vision activities with very comparable intraclass features.

Reference:

[1] M. M. Hassan, S. Huda, M. Z. Uddin, A. Almogren, and M. Alrubaian, “Human Activity Recognition from Body Sensor Data using Deep Learning,” *J. Med. Syst.*, vol. 42, no. 6, 2018, doi: 10.1007/s10916-018-0948-z.

[2] D. Nikolova, I. Vladimirov, and Z. Terneva, “Human Action Recognition for Pose-based Attention: Methods on the Framework of Image Processing and Deep Learning,” *2021 56th Int. Sci. Conf. Information, Commun. Energy Syst. Technol. ICEST 2021 - Proc.*, pp. 23–26, 2021, doi: 10.1109/ICEST52640.2021.9483503.

[3] R. Poppe, “Vision-based human motion analysis : An overview,” vol. 108, pp. 4–18, 2007, doi: 10.1016/j.cviu.2006.10.016.

[4] T. Özyer, D. S. Ak, and R. Alhajj, “Human action recognition approaches with video datasets—A survey,” *Knowledge-Based Syst.*, vol. 222, p. 106995, 2021, doi: 10.1016/j.knosys.2021.106995.

[5] R. Bodor, “Vision-Based Human Tracking and Activity Recognition.”

[6] S. Patil and K. S. Prabhushetty, “Bi-attention LSTM with CNN based multi-task human activity detection in video surveillance,” *Int. J. Eng. Trends Technol.*, vol. 69, no. 11, pp. 192–204, 2021, doi: 10.14445/22315381/IJETT-V69I11P225.

[7] S. S. Begampure and P. M. Jadhav, “Intelligent Video Analytics For Human Action Detection: A Deep Learning Approach With Transfer Learning,” *Int. J. Comput. Digit. Syst.*, vol. 11, no. 1, pp. 63–71, 2022, doi: 10.12785/ijcds/110105.

[8] D. Cavaliere, V. Loia, A. Saggese, S. Senatore, and M. Vento, “Knowledge-Based Systems A human-like description of scene events for a proper UAV-based video content analysis ☆,” *Knowledge-Based Syst.*, vol. 178, no. 2019, pp. 163–175, 2020, doi: 10.1016/j.knosys.2019.04.026.

[9] H. Yu *et al.*, “Multiple human tracking in wearable camera videos with informationless intervals,” *Pattern Recognit. Lett.*, vol. 112, pp. 104–110, 2018, doi: 10.1016/j.patrec.2018.06.003.

[10] H. Madokoro, S. Nix, H. Woo, and K. Sato, “A mini-survey and feasibility study of deep-learning-based human activity recognition from slight feature signals obtained using privacy-aware environmental sensors,” *Appl. Sci.*, vol. 11, no. 24, pp. 1–31, 2021, doi: 10.3390/app112411807.

[11] X. Yang *et al.*, “A CNN-based posture change detection for lactating sow in untrimmed depth videos,” *Comput. Electron. Agric.*, vol. 185, no. March, p. 106139, 2021, doi: 10.1016/j.compag.2021.106139.

[12] I. U. Khan, S. Afzal, and J. W. Lee, “Human activity recognition via hybrid deep learning based model,” *Sensors*, vol. 22, no. 1, 2022, doi: 10.3390/s22010323.

[13] L. Mo, F. Li, Y. Zhu, and A. Huang, “Human physical activity recognition based on computer vision with deep learning model,” *Conf. Rec. - IEEE Instrum. Meas. Technol. Conf.*, vol. 2016-July, 2016, doi: 10.1109/I2MTC.2016.7520541.

[14] P. Y. Chen and V. W. Soo, “Humor recognition using deep learning,” *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 2, pp. 113–117, 2018, doi: 10.18653/v1/n18-2018.

[15] A. A. Abed and S. A. Rahman, “Python-based Raspberry Pi for Hand Gesture Recognition Python-based Raspberry Pi for Hand Gesture Recognition,” no. September, 2017, doi: 10.5120/ijca2017915285.

[16] M. Latah, “Human action recognition using support vector machines and 3D convolutional neural networks,” *Int. J. Adv. Intell. Informatics*, vol. 3, no. 1, pp. 47–55, 2017, doi: 10.26555/ijain.v3i1.89.

[17] Z. Shi, J. A. Zhang, R. Xu, and G. Fang, “Human Activity Recognition Using Deep Learning Networks with Enhanced Channel State Information,” *2018 IEEE Globecom Work. GC Wkshps 2018 - Proc.*, 2019, doi: 10.1109/GLOCOMW.2018.8644435.

[18] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong, “Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning,” *Sensors (Switzerland)*, vol. 19, no. 7, 2019, doi: 10.3390/s19071716.

[19] O. S. Amosov, S. G. Amosova, Y. S. Ivanov, and S. V Zhiganov, “ScienceDirect ScienceDirect ScienceDirect Using the Ensemble of Deep Neural Networks for Normal and Using the Ensemble of Deep Neural Networks for Normal and Abnormal Situations Detection and Recognition in the Continuous Abnormal Situations Detection and,” *Procedia Comput. Sci.*, vol. 150, pp. 532–539, 2019, doi: 10.1016/j.procs.2019.02.089.

[20] V. Mavani, S. Raman, and K. P. Miyapuram, “Facial Expression Recognition using Visual Saliency and Deep Learning Viraj Mavani L . D . College of Engineering Shanmuganathan Raman Indian Institute of Technology Krishna P Miyapuram Indian Institute of Technology,” pp. 2783–2788, 2012.

[21] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib, “Human action recognition using transfer learning with deep representations,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, pp. 463–469, 2017, doi: 10.1109/IJCNN.2017.7965890.

[22] T. B. Moeslund, A. Hilton, and V. Kru, “A survey of advances in vision-based human motion capture and analysis,” vol. 104, pp. 90–126, 2006, doi: 10.1016/j.cviu.2006.08.002.

[23] Singh, S. ., Wable, S. ., & Kharose, P. . (2021). A Review Of E-Voting System Based on Blockchain Technology. International Journal of New Practices in Management and Engineering, 10(04), 09–13. https://doi.org/10.17762/ijnpme.v10i04.125

[24] Veeraiah, D., Mohanty, R., Kundu, S., Dhabliya, D., Tiwari, M., Jamal, S. S., & Halifa, A. (2022). Detection of malicious cloud bandwidth consumption in cloud computing using machine learning techniques. Computational Intelligence and Neuroscience, 2022 doi:10.1155/2022/4003403