Human Activity Recognition Vision Based Pose Detection

Dr. Veenu[1], Vikas*,[2], Aakash Katiyar[3]

1. Netaji Subhas University Of Technology, Dwarka Sector-3, Delhi-110078, veenu.d@rediffmail.com
2. Netaji Subhas University Of Technology, Dwarka Sector-3, Delhi-110078, vikas.mc19@nsut.ac.in
3. Netaji Subhas University Of Technology, Dwarka Sector-3, Delhi-110078, aakash.katiyar.co19@nsut.ac.in

Abstract:

Human activity detection systems in the modern era are mostly trained and operated on video feeds and picture data. This data contains modifications to the functioning and operation of data that exhibits comparable or related behaviours. Recognizing human activity is critical for both human-to-human and interpersonal interactions. Operating systems manually is time-consuming and expensive. The project's objective is to develop a minimal cost, high-speed human behaviour detection system that relies on vision and can process video and photographs to determine what is going on. Examples include surveillance, the sports industry, and healthcare. In addition to being cost-effective, the system is a time-saving, user-directed system that can be integrated into a variety of useful applications for a wide variety of activities that require cognitive processing with extraordinary precision. In addition, it can help blind people use their knowledge of their surroundings.

Keyword: Artificial intelligence (AI) strategies are being developed to distinguish human nature in UCF internet data.

1. Introduction

Existing systems were manual, requiring users to sit face-to-face with a screen to monitor and coordinate human work, which was laborious, time-consuming, overpriced, and prone to stupid mistakes and carelessness. Furthermore, some systems began to employ sensor data to identify human activity, yet they generally required the user to wear them, which limited the range of activity detection in open spaces.

The proposed methodology, unlike existing technologies, employs a view-based identification technique that does not extract complicated information from photographs and videos since it identifies using visual templates. Instead, we keep the entire raw video data as a single feature in a high dimensional space. These sample images or videos are trained with different poses and lighting conditions for recognition.

Based on his 2D rendering idea of an action video or photo sequence, he combines image sequences into his one to perform human activity detection. For classification, we use machine learning and deep learning models such as 'LRCN' and 'ConvLSTM' models to improve the accuracy of images and videos of multiple people. Focusing primarily on similar human behavioural categories such as human-object interactions, body movements only, human-human interactions, playing musical instruments, sports, etc., we test our methods on integrated datasets. Based on the results obtained, our model proves sufficient to detect human activity and, moreover, is invariant to the speed of actions performed in the above variations.

We have also extended the model to detect and calculate other information such as speed, altitude, etc. Based on the results obtained, our model is shown to be sufficient for human activity detection and invariant to the speed of actions performed in addition to the above variations.

2. Literature Review

The proposal of expressing articulated things in general, and human postures throughout specifically, as graphs of components was introduced early on and is being utilised in some cases today. Fishler and El schlager's so-called photographic strictures have become tractable and practicable thanks to distance transformation methods. As a result, several PS-based prototypes of significant utility were constructed.

However, the manageability is limited by the fact that we utilise a tree-based posture model with a basic binary capability that is irrespective of standards in this regard. As a response, the research has focused on boosting model visualisation capabilities while keeping predictability. Past attempts to do this relied on broader subdetectors [18, 1]. Alternative techniques for representing complicated sharing relationships have recently been invented.

Yang and Ramanan employ a multi-part model. Johnson and Everingham investigated a PS mixing model on a 1:1 scale. Tian et al. reported a richer, higher-order spatial connection, which was integrated into the hierarchical model. [twenty-four]. Image-dependent PS models, that can be estimated using universal learners, are another method for capturing higher-order relationships.

Techniques that explain the pose-thinking concept comprehensively have been shown to be of minimal functional significance. Mori and Malik try to discover the nearest occurrence from the collection of marked photos for each test image and communicate the shared position information. Shakhnarovich et al. suggested a similar sub adjacent structure. It does, however, employ a location dependent hash. Gkioxari et al. have suggested a semi-global classifier for component composition. This proposal performed admirably on real-world data; nevertheless, it is built on a linear classifier that is less strong than weapons, and it has only been tested on weapons.

HAR research has been carried out on a variety of datasets. Multicolour, smartphone, and complexity sensor information were used to create this large dataset. Cameras can gather data based on their location and the surrounding environment. As a result, researchers have begun to experiment with depth cameras and sensors.

In 2014, Mr. Verna Pauli proposed a model that expresses the skeleton and behaviours of the Rye group with points and curves

using 3D geometry. Luo described a dictionary-based method for skeletal data. He also proposed an action recognition model that encodes human movements and uses the construction of action diagrams represented as three-dimensional bags of points.

Nunes et al. We recognized the pose of each action window by extracting geometric and temporal data from each action window. Action classification was performed using the random forest method and the DE method was used to isolate the tree nodes. Skeletal imaging can result in noisy data due to the inability to locate body components. Chikawi et al. A combination of variables based on joints and body parts was used to detect hostile and irritating behaviours. In this work, we use the SVD method to determine the important properties.

Shahroudy et al. (2016) investigated recent methods for top-to-bottom-based human movement evaluation and achieved outstanding execution while demonstrating the feasibility of 3D portrayal for activity uniform distribution. Currently accessible profundity-based and RGB+D-based activity recognition criteria have different limitations, which included a lack of prepared tests, distinct class names, camera viewpoints, and a variety of themes. This research presents a large size dataset for human activity recognition consisting of 56,000 video tests and 4 million edges collected from 40 different people. It includes sixty different exercise courses, including daily, communal, and wellness-related activities. Furthermore, another repeated neural system structure is provided to illustrate the long-haul transient connection of all body parts' spotlights and employ them for accurate behaviour characterisation. Additionally, researchers illustrated the advantages of using deep teaching methodologies over cutting-edge approaches that incorporate cross-subject and cross-assessment parameters for the selected dataset.

Qingzhong et al. (2018) suggested a technique for detecting human activity using smartphone sensor data that is split into two categories gesture and mobile mobility. Following the extraction of characteristics, classification using machine learning techniques were used to examine human activities. Ultimately, the performance is examined that used a Convolutional Neural Network algorithm.

3. Methodology

Step-by-step process of human activity recognition Vision based pose and motion detection model follows below:

3.1. Data acquisition

We use web scraping techniques to collect data from various websites. Ethical considerations such as privacy and consent issues should be considered when scraping and collecting data.

3.2. Pre-processing

Start Processing the dataset. To begin, download the video clip from either the dataset then reformats the video frames to a set size and height to decrease calculations and divide the image pixels by 255 to normalise the data to the [0-1] range. While training the network, the convergence time is reduced.

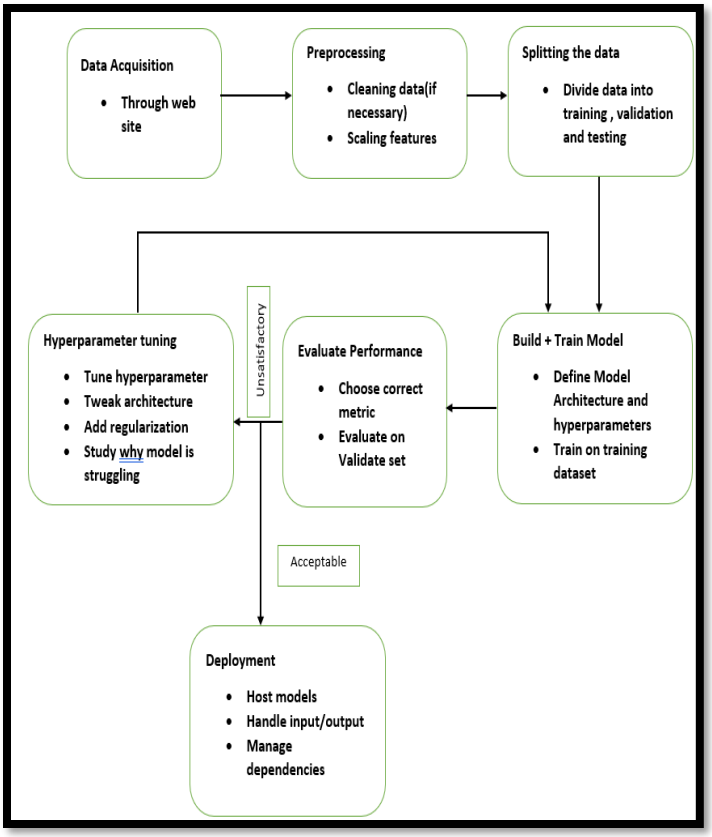

Fig. 1- Research Framework of proposed system

3.2.1. Feature extraction

The Extraction of Features phase will aid in the production of a lists comprising the scaled and normalised images of a clip whose filename is supplied as an input to it. The algorithm will accept the video frame per frame, however not all frames will be put to the list because we just require a sequential duration of frames that is uniformly distributed. Extracting features differs depending on the technique employed for human action recognition. This is necessary to lessen the ambiguity that the systems experience when trained with large feature sets. It also helps in decreasing computational burden and even enhances reliability, cutting mistake rates. The system extracts significant features that are unique to a specific activity.

3.2.2. Representation of features

The extracted characteristics must be expressed as frame (attributes), class indexes (labels), and video file locations (video files paths) so that methodologies may be implemented to them efficiently. The retrieved characteristics might be expressed as a single descriptor or as a topic model, with each collection of words representing a subject. The depiction is based on the retrieved characteristics.

3.3. Splitting the dataset

Divide the data set once it has been processed. We divided the data into two categories: training and validation. We may also establish a third holdout population known as the testing dataset. We

frequently use the phrases "validation" and "test" synonymously without doing so. We possess all the essential characteristics, including a numeric value array having all the captured frames from the video's clips and one hot encoded label, which is another Multidimensional data vector containing all class labels in one hot encoded format. As a result, we shall divide our dataset into sets for training and testing. We will additionally shuffle the dataset just before splitting for disenfranchise minorities and provide splits that accurately depict the entire distribution of the information.

3.4. Build and train model

Utilizing the training knowledge, classification techniques are utilized to develop prediction model. The proposed methodology is being utilised to evaluate videos clips for activity detection and categorization. Naive Bayes, K-Nearest Neighbours, Bayesian Decision HMM, and Feedforward Neural Networks are among the classification techniques used for activity detection. In newer techniques, neural pathways and deep learning are also utilised. Convolutional Neural Networks are used to uncover hidden patterns in a dataset, Recurrent Neural Networks are used to retrieve temporal information from time series data, and LSTMs are employed. We conduct human action recognition utilising the LRCN and ConvLSTM techniques out of all available models.

3.5. Evaluation measures

Evaluation is measured in terms of Cross Validation and Evaluation Metrics such as Accuracy and Loss of model.

3.6. Tuning hyperparameters

Statistical differences between both the learning and evaluation sets suggest overfitting, as well as the model should indeed be diminished in size or regularised such as drop out. Poor performance for both the testing and training sets indicates poor adaptation, which might also necessitate the use of a bigger model or a various learning strategy. Starting with a modest model and scaling the parameters of the model until learning and validated go independently is a frequent strategy.

3.7. Deployment and analysis

We maybe wish to apply our system in the real universe after we have developed something that. This is particularly the case when our workers and consumers utilise our systems, or when customers operate behind the scenes using commercial services and corporate tools.

4. Interface With The Client

User interaction simulator is a computational model used by computer device developers. Today's modern interfaces UIs are modern and powerful that play a significant part in the functionality of applications. Designing a UI ultimately entails not just studying rules and practice guidelines, as well as a building approach that encompasses the development of graphical frameworks and systematic vocabulary for this depiction. It uses the PyCharm IDE GUI design, which is straightforward and simple to operate to fit your regular consumption habits. This is a formal and well acknowledged idea. This is done to emphasize the public's activities.

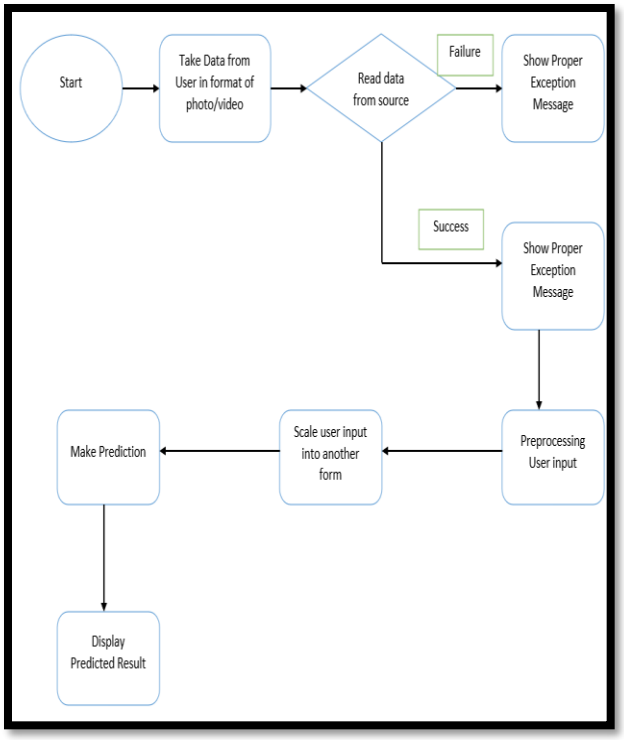

Fig 2: Interface of model

5. *Dataset*

Our human activity recognition dataset contains 101 action categories based on genuine action footage obtained from internet sites. UCF101 has the most diversity in terms of activity, with 13320 videos clips from 101 action categories, with substantial changes in camera motions, object presentation and postures, object sizing, views, backdrop distraction, illumination changes, and much more. Yes, it richest and most complicated dataset till present. Because the majority of accessible activity recognition datasets are impractical and performed by performers, this dataset may drive additional activity recognition research by learning and investigating new demand action classes. is designed for the 101 activity class clips are divided into 25 subgroups, with each grouping consisting of approximately 7 action clips. Videos within the same section may have similar qualities. Behaviours categories can be divided into five types. Movement from these categories, we use human activity videos of basketball, cycling, bowling, push ups, typing, walking with dogs, and writing on board to train models.

6. *Model description*

Action recognition is one of the hottest research areas these days. Activity classification is essentially a time series problem. Time series classification is a type of supervised machine learning. It is used to predict future values from past data using statistical techniques and can be used to predict and exchange sensor data.

The capacity to perceive and identify an agent's behaviours or intentions is a subdomain of vision-related applications. Any item or thing that has a purpose and executes an action can be considered an agent. A single agent can conduct an action, or a collection of agents can perform or modify an action. Humans are a kind of certain agents, and the recognition of human behaviour is known as "Human Activity Recognition".

The input of the HAR model is the raw dataset in form of video or photo readings for detecting human activity, and the output is the prediction of the user's motion activity. In a vision-based approach, the system uses cameras to collect data and identify activity. The resulting data is processed by an algorithm to generate a series of numbers representing human activity in machine language. Based on movements, algorithms generate predictions or insights for further analysis.

We are implementing the Human Activity Recognition Vision-based Pose Detection Project by using two approaches:

6.1. ConvLSTM

The first technique is implemented using a mixture of ConvLSTM cells. A ConvLSTM cell is an LSTM network variation that contains convolutional functions inside the circuit. It's an LSTM having convolution embedded into its structure that enables you to detect geographical data properties while taking temporal interactions into account. This method efficiently identifies geometric correlation inside each frame as well as temporal correlations between distinct frames in video categorization. Because of its convolutional design, ConvLSTM can take 3-dimensional inputs (width, height, number of channels), but naïve LSTMs only accept 1-dimensional inputs, making LSTM incompatible with modelling spatio-temporal data.

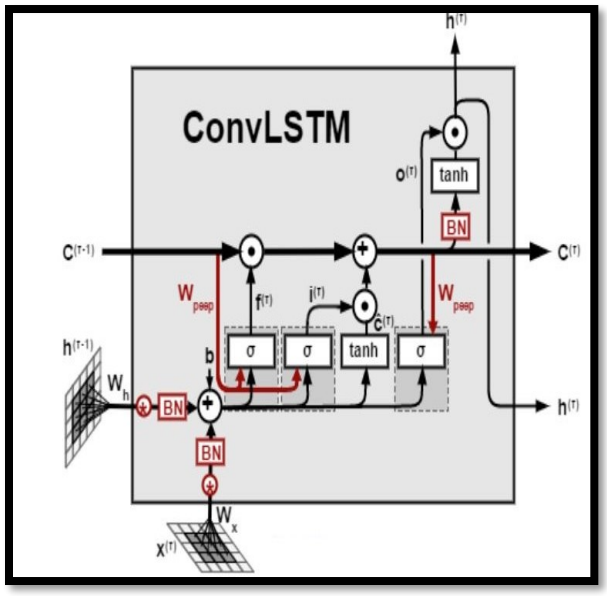

Fig 3: Architecture of ConvLSTM

Keras ConvLSTM2D repetitive layers are used to construct the classifier. The ConvLSTM2D layer is also in full control of the whole of the quantity of filters and kernel sizes needed to perform the convolution process. Eventually, the layer's outcome is smoothed and fed into a SoftMax-enabled dense layer. This will generate the chances for each action type. Use drop - outs layers to discourage the classifier against overfitting the information and MaxPooling3D layers to minimise frames and eliminate wasteful calculations. The design is straightforward, with only a few learnable parameters. This is a result of the fact that we are merely interacting with a tiny portion of the data set, which does not necessitate a huge framework. We also provide a premature stop trigger to build a predictive model as well as to begin training immediately after model compilation.

6.2. LRCN

The second methodology incorporates convolutional and LSTM layers together into unified framework to execute the LRCN methodology. Another strategy that is comparable is to utilise independently learned CNN and LSTM networks. CNN models may be employed to retrieve spatial information from consecutive frames. We can employ pre-trained algorithms which can be perfectly alright to your situation for all of this objective. The CNN characteristics may then be used by the LSTM model to forecast actions in the clip. However, in this case, we use the Long-Term Recurrent Convolutional Network (LRCN), that also integrates CNN as well as LSTM layers together into unified framework. To represent the time series, a deep neural network is utilised to retrieve spatial characteristics from frames, and the collected characteristics are supplied to the LSTM layer through each time interval. In this manner, the network can be trained spatiotemporal properties automatically in end-to-end learning and develops a reliable prediction.

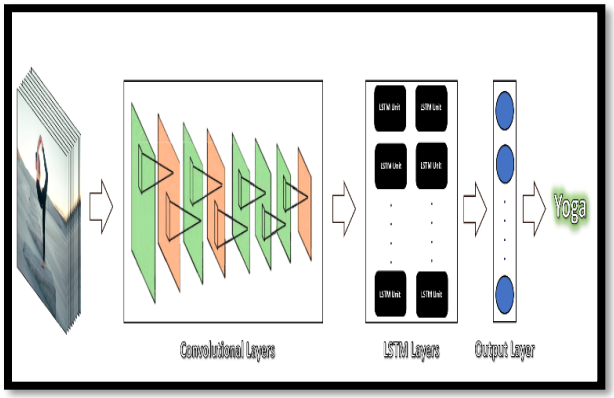

Fig 4: LRCN working

We utilize a time-splitting Conv2D layer, accompanied with MaxPooling2D as well as Drop - outs layers, to create the LRCN framework. The Conv2D layer's characteristics are flattened only with Flatten layer before being transferred to the LSTM layer. The outcome of such LSTM layer is then used by a SoftMax-enabled dense layer to forecast the actions to be executed.

One of the biggest challenges in human activity detection is estimating human pose. This has traditionally been achieved using hand-crafted models that require careful initialization and parameter estimation, but recent advances in deep learning have made it possible to use neural networks.

7. Result

Recognition of human activities software estimate activity in videos clips with amazing precision and photos having middling precision. Here is a glimpse of the models' collective accuracy and loss.

7.1. Accuracy and loss of ConvLSTM

Fig 5: Total loss v/s total validation loss

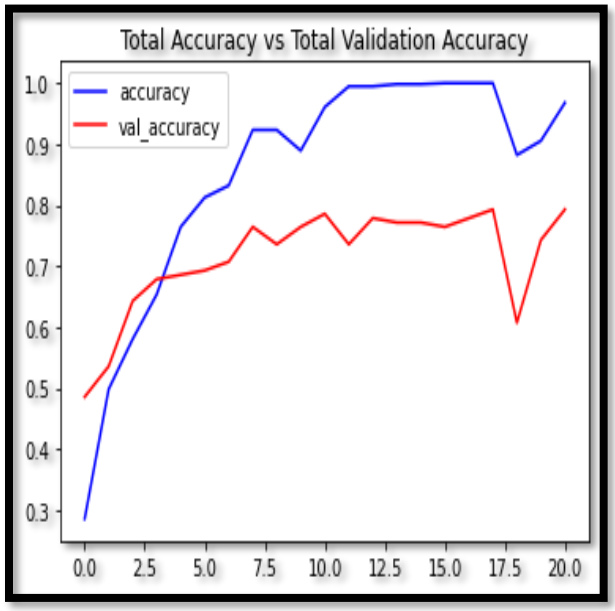

Fig 6: Total accuracy v/s total validation accuracy

7.2. Accuracy and loss of LRCN

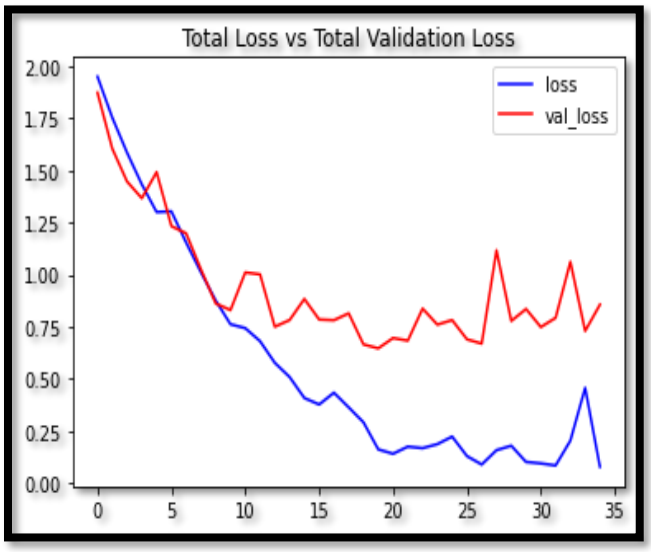

Fig 7: Total loss v/s total validation loss

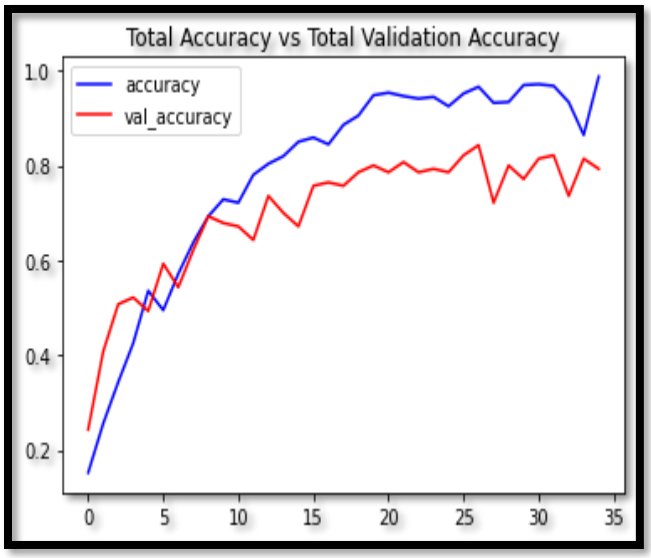

Fig 8: Total accuracy v/s total validation accuracy

In this, we can see ConvLSTM has the total Validation Accuracy with 0.752 and total Accuracy with 0.936 whereas LRCN Neural Network gave total Validation Accuracy at 0.826 and total Accuracy at 0.965. Fig. 6 shows the comparison of validation accuracy and accuracy for the ConvLSTM neural network, and Fig. 5 shows the comparison of validation loss and loss for the ConvLSTM neural network. Fig. 8 shows the comparison of validation accuracy and accuracy for the LRCN neural network, and Fig. 7 shows the comparison of validation loss and loss for the LRCN neural network. From Fig. 6 and Fig. 8, which clearly indicate the LRCN neural network gave 96.55% accuracy for the testing data set, this is better than the former approach.

8. Limitation

Human behaviour identification has significant ramifications in a variety of industries, including safety, medical, visual monitoring, and numerous others. Despite its broad use, the suggested system technology has certain limitations.

Whenever numerous performers are involved within the same image, that becomes difficult to identify every one of their activities. Specific performers may be covered or just partially viewable in some settings.

Overall effectiveness of activity identification reduces due to the varying movement patterns of various participants at distinct intervals.

9. Conclusion and prospective use

Currently nights, computer vision is a critical topic., so system like human activity recognition systems solve a wide range of applications, such as monitoring, and assisting the elderly and the blind very useful and effective for It can be used not only by end users, but also by various organizations to reduce the workload of their employees. This model gives good results for video streams and works well for image data. Activity recognition systems are very important today because of the convenience and problems they provide and solve. Surveillance the need for activity detection such as video segmentation are in high demand where this system can help a lot. The system can be integrated into mobile apps to further assist the elderly and visually impaired. This is a moderate, time especially solution which is equally susceptible to human mistake. This technology serves as the foundation for a variety of additional activity monitoring applications. As a result, this approach is extremely advantageous for the both people and businesses for broad or specific goals.

This project has great implications for the future. First, the video detection code can be further refined through transfer learning to use much larger datasets to further improve the accuracy of the model. In addition, build web and mobile apps that can call these Python scripts via API calls to provide activity detection for mobile users, enable elderly or blind people to control their environment, etc. We can also help you understand and interact with them. Application of activity detection system.

10. Acknowledgements

The Department of Computer Engineering at Netaji Subhas University of Technology in Delhi provided support for this research. All claims of fact, opinion, or conclusion expressed herein are the authors' own and should not be interpreted as representing the sponsors' official views or policies.

REFERENCES

Gourav Singh (2021). *Understanding Architecture of LSTM Article. Data Scientist Blogathon*, Analytics Vidhya

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge L. Reyes-Ortiz (2012*). Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine. Springer* International Workshop on Ambient Assisted Living. Lecture notes in Computer Science. Vol (7657)

Qingzhong Liu, Zhaoxian Zhou, Sarbagya Ratna Shakya, Prathyusha Uduthalapally, Mengyu Qiao, and Andrew H. Sung (2018). *Smartphone Sensor-Based Activity Recognition by Using Machine Learning and Deep Learning Algorithms*, International Journal of Machine Learning and Computing, Vol. 8, No. 2.

Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang (2016), *Spatio-temporal LSTM with Trust Gates for 3D Human Action Recognition*, European Conference on Computer Vision, pp. 816–833, Vol.

From Wikipedia, *Residual neural network Article*

Manish Gupta (2020), *Understanding ResNet and its Variants* Article, Published in Towards Data Science

Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell (2016), *Long-term Recurrent Convolutional Networks for Visual Recognition and Description,* Originally presented at CVPR 2015 (oral). Updated version (accepted as a TPAMI journal article)

J. K. Aggarwal and M. S. Ryoo, *Human activity analysis: A review*, ACM Comput. Surv., vol. 43, no. 3, 2011, doi: 10.1145/1922649.1922653

S. A. R. Abu - Bakar, *Advances in human action recognition: An updated survey*, IET Image Process., vol. 13, no. 13, pp. 2381–2394, 2019, doi: 10.1049/ietipr.2019.0350

A. G. D'Sa and B. G. Prasad, *A survey on vision based activity recognition, its applications and challenges*, 2019 2nd Int. Conf. Adv. Comput. Commun. Paradig. ICACCP 2019, pp. 1–8, 2019, doi: 10. 1109/ICACCP.2019.8882896