IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HUMAN ACTIVITY POSE PREDICTION USING LSTM ALGORITHM

PADALA DIVYA SANTHI

PG Scholar, Department of Computer Science,

SVKP & Dr K S Raju Arts & Science College,

Penugonda, W.G.Dt., A.P, India

CHIRAPARAPU SRINIVASA RAO*

Associate Professor in Compute Science,

SVKP & Dr K S Raju Arts & Science College,

Penugonda, W.G.Dt., A.P, India

ABSTRACT

Human motion modelling is a classical problem at the intersection of graphics and computer vision, with applications spanning human-computer interaction, motion synthesis, and motion prediction for virtual and augmented reality. Following the success of deep learning methods in several computer vision tasks, recent work has focused on using deep recurrent neural networks (RNNs) to model human motion, with the goal of learning time-dependent representations that perform tasks such as short-term motion prediction and long-term human motion synthesis. In this project LSTM algorithm is used to train user activity like sitting, standing, dancing and based on live input video file motion prediction of user is done and lines are drawn on person based on his position of activity.

KEYWORDS:

Human Activity Recognition, Convolution, Long Short-term Memory, Mobile Sensors.

INTRODUCTION

1.1 Introduction

In recent years, human activity recognition (HAR) and classification have gained momentum in both industry and academic research due to a vast number of applications associated with them. One area in particular where this research has huge interest is smart homes and the Internet of Things [14,19]. Other areas include crowd counting, health and elderly care, in particular fall detection, [20], [3]. Fall detection has been a popular area of

research to enable more independent living for both the elderly and disabled within their own accommodation, but also within environments where cameras cannot be used due to data protection. There are two main approaches used for HAR: invasive and non-invasive. Invasive HAR involves wearing sensors to track humans to create a rich dataset for models to learn from, while non-invasive HAR allows humans to be monitored without any attached devices [21]. One way to do this is using WiFi signals, which are widely available in most buildings. In HAR, the main activities in the classification task are sitting, standing, walking, lying down, falling down and human absence. All of these activities are of interest in the area of smart homes, while the falling down activity is of particular interest in health and elderly care, where cameras cannot be installed in private rooms but there is a need to monitor patients. This non-invasive, data sensitive method to alert staff to a patient falling is of great interest to the industry. The idea behind HAR using WiFi signals is that the human body will affect the signal by reflection, and that different activities will show distinct characteristics. Initially most research in this area was carried out using the received signal strength (RSS), due to its accessibility [1]. With the development of the WiFi Network Interface Card, the data-rich Channel State Information (CSI) provides fine-grained information on multiple subcarriers [5]. The CSI carries the amplitude and phase for each subcarrier in orthogonal frequency-division multiplexing. By averaging the CSI across its subcarriers one can calculate the corresponding RSS, thus showing how much more information is carried on the CSI data compared to the RSS. The CSI data is typically converted into a spectrograph image, with the axes being time and channel, and the colour representing signal intensity. Since deep learning (DL) is state-of-the-art at image recognition, it is a very suitable choice for this application. Hence, we propose a DL method to classify human activities using CSI WiFI data. There has been much recent work on HAR models, where feature engineering is required and the classification part of the tesk is preformed by traditional methods such as Support Vector Machines (SVM) or Decision Trees. Extracting features in this manner may result in fewer informative features and, because of the sequential nature of the data, subsets of features which are relevant to the classification task might be ignored. To overcome these issues, our method automatically learns the most discriminative features to distinguish each activity from each other. The main part of the proposed model is the TCN, consisting of a 1D fully-convolutional network and causal convolutions, which are convolutions where

an output at time t is convolved only with elements from time t and earlier in the previous layer [2].

2. LITERATURE SURVEY

Zou et al. [24] introduced a model called Auto-Encoder Long-term Recurrent Convolution Network (AE-LRCN), which was a DL approach to HAR. An autoencoder was used for representation learning and to remove inherent noise. The raw CSI time window was transformed into a latent space of 256 features, whereas our proposed method used a CAE not only to remove noise, but to compress the CSI time window into a latent space of 12 features. Next Zou et al. [24] used a convolutional network for feature extraction, which had 2 convolution layers followed by 2 fully connected layers. The proposed model does not require this step as the autoencoder had already performed aggressive feature selection. Finally, for sequential learning

Zou et al. [24] implemented the pop- portant features from accelerometer raw data [10, 12, 16]. Yousefi et al. [22] pre-processed channel state information CSI by means of principle component analysis to de-noise the signal, and then used a short-time Fourier transform to extract features. These features were then used as inputs to random forest (RF) and hidden Markov model (HMM) classifiers. The results were compared to a DL approach using a LSTM, which did not require de-noising or feature extraction as these are performed within the model. The RF and HMM models achieved 64.67% and 73.33% respectively. The LSTM model scored 90.5% accuracy: over 17% better than the HMM model and approximately 26% better than the RF. These results show that DL models can outdo classically methods, though the author noted that LSTMs were much slower to train. This is where our proposed model helps as it is significantly faster to train, as mentioned below.

Zou et al. [24] introduced a model called Auto-Encoder Long-term Recurrent Convolution Network (AE-LRCN), which was a DL approach to HAR. An autoencoder was used for representation learning and to remove inherent noise. The raw CSI time window was transformed into a latent space of 256 features, whereas our proposed method used a CAE not only to remove noise, but to compress the CSI time window into a latent space of 12 features. Next Zou et al. [24] used a convolutional network for feature extraction, which had 2 convolution layers followed by 2 fully connected layers. The proposed model does not require this step as the autoencoder had already performed

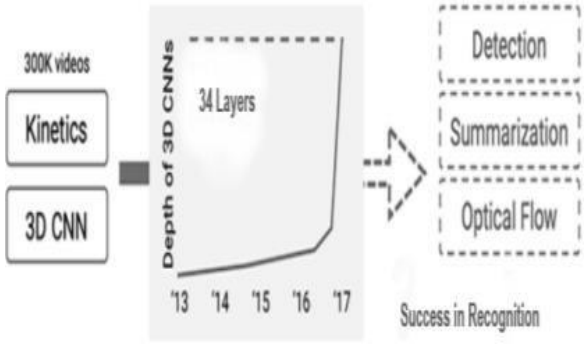
aggressive feature selection. Finally, for sequential learning

Zou et al. [24] implemented the pop- ular LSTM, which has been shown to perform very well on this type of data. We introduce the TCN model, which is new to the area of HAR using CSI data, to learn the sequential nature of the data. We show that both the TCN and LSTM achieve state-of-the-art results in HAR, but the new proposed method is much more efficient. Wang et al. [18] introduced a DL-based channel selective activity recognition system called CSAR. This method requires considerable pre-processing, starting with channel quality evaluation and selection. They select the channels with an amplitude over a threshold, and neglect the others under the assumption that they are uninformative. Next they use channel hopping, where CSAR circularly hops through these selected channels, combining adjacent channels into an extended channel with higher bandwidth.

Wang et al. [18] denoise the data by using a low-pass filter with a cut-off frequency of 100Hz, and PCA for data reduction and de-noising. Finally, the DL model implemented is the LSTM. In our proposed model the CAE denoises the signal, while the TCN — which is significantly faster than the LSTM — is used to learn the sequential nature of the CSI data. Wang et al. [18] used similar activities to our work, achieving on average over 95% accuracy, which compares well to our results.

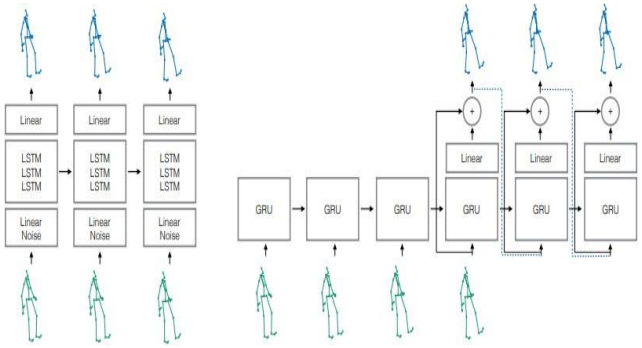
3.Methodology:

The two crucial steps for the implementation are training and recognition. Stochastic gradient descent is used. Training samples are generated from the training data. To proceed next, a temporal position in the video is selected for the generation of training samples. Then around the selected position, a sixteen-frame clip is generated. Start looping around the video as much as required if it is less than sixteen-frames. Next, a spatial position and spatial scale is selected as per requirement. Then follow the corner cropping strategy and crop four corners to 112 X 112 pixels. A weight decay of 0.001 is included in the training parameter. Learning rate is started from 0.1 and after validation loss saturates, reduce it to 0.01.



For the recognition, looping over each frame from 0 to sample duration is done. A frame is grabbed by reading the video and if the image is grabbed successfully then return true. Then the frame is resized to
400 pixels and added to the frame list. Next, the frame is resized again so that the images are of the same size. Once the frame array is completely filled then creation of the blob will begin. A blob of input frames is constructed which is passed through the network. The image is preprocessed first to obtain the correct prediction. Subtraction of mean value and scaling by using a factor of 0.1 is done to preprocess the image. This helps to make the image less sensitive to background and lighting conditions. All this is done with the help of OpenCV's deep neural network module. This blob is created to have images with the same spatial dimensions. A forward pass is applied after this and the system grabs a label with the correct prediction value. Kinetic dataset is used to train the model. This dataset consists of
400 classes of human activity including more than 650,000 trimmed videos which lasts around 10 seconds. The videos are resized without any change in the aspect ratio. The dataset contains a huge range of activities including applauding, archery, bartending, crying, handshaking, and many more. This dataset includes frames that relate to target action. The absence of noise and unrelated frames makes this dataset the most suitable for training. The number of training, validation and test sets are approximately 580000, 30000, 40000 respectively. Training the Resnet-34 model on kinetic dataset does not result in overfitting. The result of this experiment can be very important for future progress in the field of computer vision.

4. Architecture:



5. OVERVIEW OF THE SYSTEM

5.1 Existing System

Traditional approaches have typically imposed expert knowledge about motion in their systems in the form of Markovian assumptions, smoothness, or low dimensional embeddings.

Disadvantages:

- Traditional approaches have typically imposed expert knowledge about motion in their systems in the form of Markovian assumptions smoothness, or low dimensional embeddings.

5.2 Proposed System:

In proposed system LSTM algorithm is used where trained model is used for prediction of human activity recognition like sitting, standing, running, dancing.

Etc. Prediction will be as lines are drawn on human body based on his position.

Advantages:

• We have demonstrated that previous work on human motion modelling using deep LSTM has harshly neglected the important task of short-term motion prediction, as we have shown that a zero-velocity prediction is a simple but hardtop-beat baseline that largely outperforms the state of the art.

we have developed a sequence to-sequence architecture with residual connections which, when trained on a sample-based loss, outperforms previous work..

5.3 System Modules

Dataset collection:

This was a 6-class classification problem. The classes were sit, stand, walk, lay, fall and empty, which were chosen to represent the standard range a person would carry out on a daily basis. The lay class was based on lying down on a desk, used to simulate a person lying down on a bed. The fall class simulated a person falling and remaining on the ground. These two activities were purposely picked to show how this DL method could be used in a real-world situation. Two sets of data were collected. The first, SetA, collected all the samples of each activity on the same day; the second, SetB, consisted of 3 samples of all the activities collected on a single day over a period of 7 different days. SetA had 20 samples over 180 seconds of each activity; the first and last 60 seconds was of the empty room, while the center 60 seconds was of the activity. SetB was collected differently, as 3 random sequences of each activity were performed for 60 seconds each, resulting in 1080 seconds of CSI data for each day. SetB had 126 samples in total, and was used to test the temporal robustness of the setup.

pre-processing:

The CSI data is received in complex form, the real number representing the amplitude and the imaginary number representing the phase. As we are dealing with only a single transmitter and receiver setup, the phase is of little value and is discarded. An array of the amplitude of the 90 subcarriers, 30 within each frequency band, is saved to an array. This array consisted of 20 seconds of an activity, where the CSI data was sampled 5 times per second. This 90 × 100 array was then split into 4-second windows with a 50% overlap. Each 20-second activity array was transformed into a sequence of nine 90 × 20-time windows, which formed the dataset that the CAE was trained and validated on. The usual 80-10-10 training, validation and testing split was used to

ensure no leakage into the training phases of the CAE or TCN.

Initialize algorithm:

In this stage LSTM algorithm is used and training data is fit in to algorithm.

Save Model:

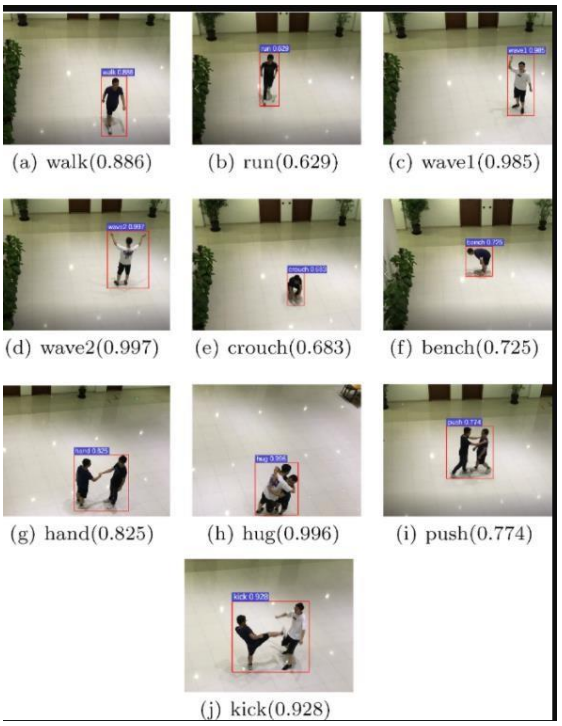
After training data is saved as .h5 model which is used for predicting purpose in future.

Predict:

In this stage we take recorded video as input and predict type of activity.

6. RESULTS

Detecting action



Action detection result

7. CONCLUSION

In this paper, we presented a novel LSTM, based on our experimental results, to accurately help solve the HAR problem. We design a Convolutional Neural Network Model to helps to remove unwanted noise in the pre-processed CSI data, while compressing it through a bottleneck embedding layer. This compressed latent layer, vector size 12, was used to build a sequential model for activity classification. By embedding the CSI window time steps onto a lower dimensionality, it greatly reduces the problem complexity, therefore allowing for real-world applicational purpose. The main contributions of this paper are that the algorithm is computational more

efficient, achieves state-of-the-art results, and is more robust than ANN on temporal variance in the activity data.

Future Enhancements:

In future work we plan to explore the use of transfer-learning to distil knowledge learnt in one room onto testing in a new environment. Other future work may include extending from single person to multi-person classification.

REFERENCES

[1] A. Chaudhary, J.L. Raheja, K. Das, S. Raheja, "A Survey on Hand Gesture Recognition in context of Soft Computing", Published as Book Chapter in "Advanced Computing" CCIS, Springer Berlin Heidelberg, Vol. 133, 2011, pp. 46-55

[2] A. Chaudhary, J.L. Raheja, K. Das, "A Vision based Real Time System to Control Remote Robotic hand Fingers", In proceedings of the IEEE International Conference on Computer Control and Automation, South Korea, 1-3 May, 2011, pp. 118-122

[3] J.L. Raheja, A. Chaudhary, S. Maheshwari, "Automatic Gesture Pointing Location Detection", Optik: International Journal for Light and Electron Optics, Elsevier, Vol. 125, Issue 3, 2014, pp. 993-996.

[4] A. Chaudhary, K. Vatwani, T. Agrawal, J.L. Raheja, "A Vision-Based Method to Find Fingertips in a Closed Hand", Journal of Information Processing Systems, Korea, Vol. 8, No. 3, Sep 2012, pp. 399-408.

[5] A. Chaudhary, J.L. Raheja, K. Das, S. Raheja, "Fingers' Angle Calculation using Level-Set Method", Published as Book Chapter in "Computer Vision and Image Processing in Intelligent Systems and Multimedia Technologies", IGI,

USA, April 2014, pp.191-202

[6] D. Sturman, D. Zeltzer, "A survey of glove-based input", IEEE Transactions on Computer Graphics and Applications, Vol. 14, No. 1, Jan. 1994, pp. 30-39.

[7] T.S. Huang, and V.I. Pavlovic, "Hand gesture modeling, analysis and synthesis", Proceedings of international workshop on automatic face and gesture recognition, 1995, pp.73-79.

[8] O.Y. Cho, and et al. "A hand gestue recognition system for interactive virtual environment", IEEK, 36-s(4), 1999, pp. 70-82.

[9] M. Sonka, V. Hlavac, R. Boyle, "Image processing, analysis, and machine vision"2014, Cengage Learning.

ABOUT AUTHORS:

PADALA DIVYA SANTHI is currently pursuing MCA in SVKP & Dr K S Raju Arts & Science College, affiliated to Adikavi Nannaya University, Rajamahendravaram. Her research interests include Data Structures Web Technologies, Operating Systems and Artificial Intelligence.

CH.SRINIVASA RAO is a Research Scholar in the Department of Computer Science & Engineering at Acharya Nagarjuna University, Guntur, A.P, India. He is working as Associate Professor in SVKP & Dr K S Raju Arts&Science College, Penugonda, A.P. He received Masters degree in Computer Applications from Andhra University and Computer Science & Engineering from Jawaharlal Nehru Technological University, Kakinada, India. He Qualified in UGC NET and AP SET. His research interests include Data Mining and Data Science.