EasyChair Preprint

№ 2297

Human activity recognition with openpose and Long Short-Term Memory on real time images

Chinmay Sawant

January 2, 2020

Human activity recognition with openpose and Long Short-Term Memory on real time images

Chinmay Sawant
Independent Researcher
Pune, Maharashtra – 411057, INDIA
chinmayssawant44@gmail.com

***Abstract*— Human Activity Recognition(HAR) is a broad field of study aims to classify time series activities. These activities can include normal body movements such as standing up, sitting down, jumping, walking etc. Whenever activity is performed by a person, it usually few seconds. Image classification algorithms fails to classify such stream of images into activity class. Existing approaches rely on the sensor data recorded by accelerometer, smart phones, or some harnessing devices to recognize the movements. Such data is challenging and expensive to collect and requires multiple sensors, custom hardware and software.**

In this paper, I have described a systematic method to recognize human activities in real time using Openpose and Long short-term memory networks. This approach is based on the images that are captured in real time by connecting the camera and fetching the timed screenshots. Openpose is an open source, real time project to jointly detect human body with hands, face, facial expressions, legs on single image. Output of body features are split into sub-sequence called window using sliding window approach. Long Short Term Memory(LSTM) is a Recurrent Neural Network(RNN) different from feed forward Neural Networks. It has feedback connections so it can process single data points as well as sequence of data. LSTM is suitable for this scenario and provides improved results. It efficiently learns the key point features and returns an activity class. This system detects Real time human activities such as waving 1 hand, waving 2 hands, jumping_jacks, boxing, jumping, clapping.

Keywords*— *Convolutional Neural Networks, openpose, LSTM, Image processing, sliding window approach

I. Introduction

Human Activity Recognition system continuously monitor human activities which can be helpful in surveillance, health care, anomalous behaviour detections, personal identity, knowing psychological state, elderly care. The activities can be human to human, independent or human to object interactions and can be monitored using video surveillance, wearable sensors, human to system interactions.

In spite of the advancements done in the field of Human Activity Recognition[16][17], it still faces major dependencies such as data input from wearable sensors and their installations, accuracy and inconvenience[1][4][5]. Data input is done via installing sensors throughout the body. Smart phone, smart watches collects the data with accelerometers, gyroscope[20]. The interpretations of activities is done by analyzing pattern over the time period. Activities such as sleeping can be detected by interpreting negligible movements in the data collected by accelerometers. System based on video surveillance identifies the person, creates a unique identity and try to identify the activity. The development of such a system with low error is challenging because of the background noise, subject occlusions, low light conditions, and the difficulties faced in tracking a person.

The presented technique of classifying human activities is promising and solves the issues mentioned with the above approaches. The data is gathered with the help of connected cameras and the algorithms perform real time. This eliminates the need to attach the wearable sensors and makes easy, convenient to use in any situation. Openpose[2] extracts the body key point features in the skeleton form handling cases such as background noise, low light distorted conditions. This makes the given algorithm robust in nature.

In this paper, following main contribution are made:

- I present the real time system to classify human activities with the help of connected cameras.
- Elimination of multiple wearable sensors and data acquisition thereby improving the scalability and easy to use in every field.
- A robust approach with the help of robotics research i.e. openpose and deep learning i.e. Long Short Term Memory.

II. DATA SET DESCRIPTION

The dataset is gathered from Berkeley MHAD[3]. It is a multimodal human action database for 11 different activities performed by 12 subjects. The activities are recorded from 10 different views with audio recordings. Each subject performs one activity 5 times which results in 82 minutes of recording. The dataset is freely available for research purposes.

The approach has been implemented with below 6 different activities.

- Waving 1 hand
- Waving 2 hands
- Jumping jacks
- Boxing
- Jumping
- Clapping

These are available at Github repository: https://github.com/chinmayembedded/Human-Activity-Recognition

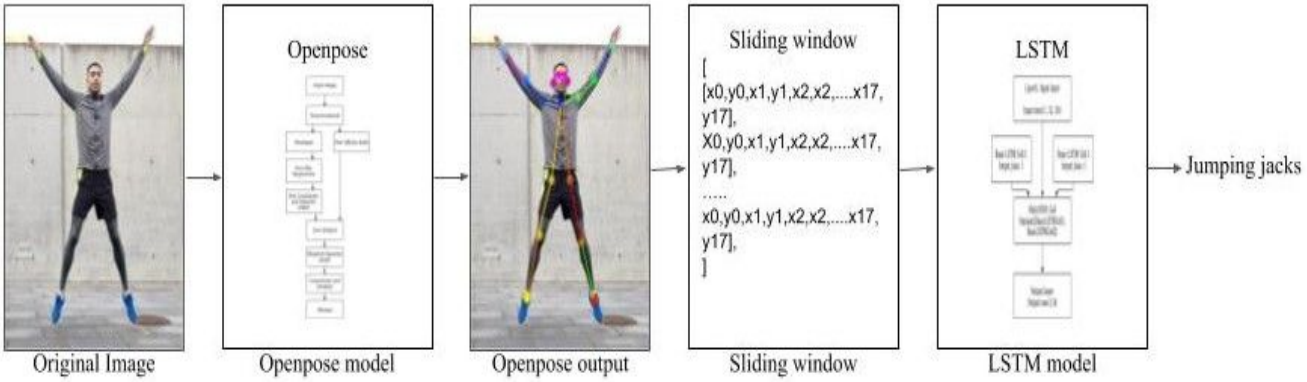

Figure 1: represents the sequential pipeline model for detecting real time human activities. Input image can be captured from video, webcam or any other cameras. It is passed to openpose model

III. APPROACH

Figure 1 presents the process flow of detecting real time human activities. Original image is obtained with the help of cameras, pre recorded videos.

Openpose model detects the 18 key points if a human body is present as renamed as [x0, y0, x1, y1, x2, y2, x3, y3, ….. ,x17, y17].

Current implementation processes 10-11 frames per second. Assumption made is to perform any activity a person takes approximately 3 seconds.

Sliding window approach[8][9] works in real time and forms an array of 32 sequential openpose outputs. It is shown in figure with sliding window label. It is then passed as an input to LSTM[6][7][21] model. It outputs the score for each label. The label with the highest score is the activity performed.

A. Image processing

Input stream is read with the help of OpenCV library functions. Per second, 10 equally spaced frames are selected. Each pixel values from an image ranges from 0 to 255. Image pixel intensities are converted from [0,255] to [-1,1].

$$Image = (Image/255) - 1$$

B. Openpose

Openpose(figure 2) is an open source human pose estimation library. It detects the human body key points, facial expression and positions, hand and foot key point extraction. The pretrained openpose model can give 15, 18, and 25 key descriptors for a human body. Openpose model is trained with COCO dataset to extract 18 body key point coordinates.

Input images are read from pre recorded videos or cameras. Openpose uses a neural network which returns a tensor containing 57 matrices. It outputs heatmaps and Part Affinity Fields. The output tensor is a concatenation of these 2 fields.

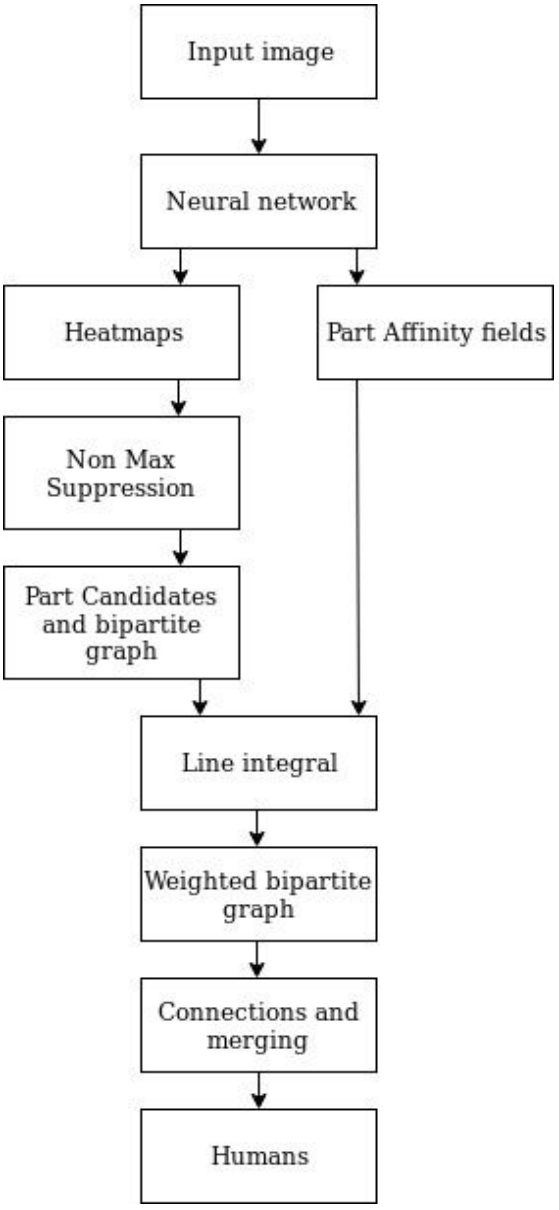

Figure 2. Openpose architecture

- Heatmaps

Each heatmap stores a matrix which contains the confidence that a pixel contains body part. The skeleton shows 18 heatmaps are associated with each one of the body parts. The location of each body part is extracted with this 18 matrices.

- Part Affinity Fields

Part Affinity Fields are matrices which contains position and orientation of pairs. For every keypoint there is a PAF in *x* direction and one in the *y* direction. There are 38 such matrices which forms the skeleton of a person.

Next layer is Non Max Suppression layer(NMS). It gives the certainty for the heatmap confidence obtained for each body part. In other words, we need to extract local maximums out of a function.

As we have found out the coordinates of the body parts, we need to join them to form skeletons. Bipartite graph connects the neck and body candidates. The vertices are body parts and association between them is represented by connection candidates.

C. Sliding window approach

Sliding window approach[14][15] is a problem solving technique for sequential data in array/collections. The fixed sized window or a fixed sized subset is selected from a collection or an array. Right shift by one is performed to select the next window.

In real time, the openpose output of 32 sequential frames forms a window and given input to a trained LSTM model which is defined in next architecture. For every 32 frames window the model learns features and outputs an activity performed. If there isn't a person present in the frame the array of zero is formed and passed to a network resulting in no activity performed.

D. LSTM

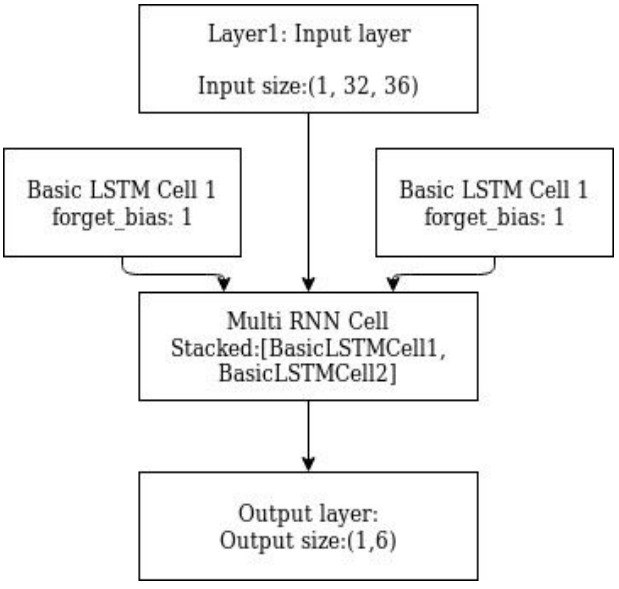

Figure 3. LSTM architecture

Long Short Term Memory(LSTM)[10][11] as shown in figure 2 is a Recurrent Neural Network(RNN)[12][13]. It has a forget gate which remembers the past states the model has learnt gives output to the current state.

Activity recognition is done by training on the outputs of openpose. For the mentioned 6 activities, 18 coordinate body key points are fetched with the help of openpose. The output of 32 sequential frames is considered as one activity So the training data contains batches of 32 sequential frames and an activity label associated with it There are 7426 batches used for for 6 activities.

IV. RESULTS

I have implemented tensorflow openpose library for human body keypoint extraction. Input data to LSTM model is an output of openpose obtained on Berkeley MHAD dataset. For the trained activities mentioned above in dataset description, human skeleton points used in a trained file and a label is mentioned for 32 sequential data inputs. If the person is not present/ performs any activity on which model is not trained is detected as no activity detected. Zero value array of human body skeleton is formed for no activities.

Table 1. Experiment results

Total number of images	237632
Total number of activities	6
Number of batches	7426
Training accuracy LSTM	**92.56%**
Testing accuracy LSTM	**87.12%**

As shown in table 1, the training of Long Short Term Memory network gives 91.56% accuracy. The data is split into 80% as training data and 20% as testing data set. Model is trained for 200000 iterations and results are obtained with cross validation. The accuracy on testing dataset is 87.12%.

The final model is integrated with openpose and real time camera streaming mechanism. The 32 sequential output array is formed by sliding window approach. The model performs accurately when whole human body is present in the image. There are cases when partial body is present in the image, similar activities such as boxing and clapping can be detected inaccurately. Distinctive activity performed for each of those results in correct activity recognition. The real time scenario of such kind can be solved by taking the maximum over previous outputs for certain period.

The results obtained are outstanding with the start of a work. The LSTM network is almost always able to identify the human activity type. I am amazed by the results obtained with activities such as jumping jacks and jumping, boxing and clapping. These activities are similar because of the same body posture and similar input features.

Figure 4: Results with the help of Openpose and Long Short Term Memory. Note to make, results though appear on images in red font, it's the activity learnt for 32 frames captured at 10 fps.

The real time system performs at 10-12 fps on medium resolution images acquired by webcam and 5-6 fps for high resolution images. The results are shown in figure 4.

V. FUTURE SCOPE

Human Activity Recognition has wide range of applications in multiple domains. It becomes easy to use when working with the images rather than installing sensors all over the body. Existing system can be integrated with hand and facial gesture recognition. By identifying unique person identity in an image, it can recognize multiple human activities. It is developed with tensorflow python Application Program Interface. Caffe installation can improve this result upto 25-30 frames per second.

VI. CONCLUSION

Image based approach has been presented for Human Activity Recognition. The aim is to convert sensor based data capture to flexible, easy to use and install camera based system. This model can be an enabler for students/ researchers to further improve and achieve wide real time use cases. I explained the approach and constraints in the existing system.

Presently, the implementation is open sourced on Github and will be updated with features and improvements.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Weil, Yaser Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," arXiv:1812.08008

[3] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, Ruzena Bajcsy, "Berkeley MHAD: A comprehensive Multimodal Human Action Database", 2013 IEEE Workshop on Applications of Computer Vision (WACV)

[4] Itishree Mandal, S L Happy, Dipti Prakash Behera, Aurobinda Routray , "A framework for human activity recognition based on accelerometer data", IEEE 2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence)

[5] Vu Ngoc Thanh Sang, Nguyen Duc Thang, Vo Van Toi, Nguyen Duc Hoang, Truong Quang Dang Khoa, "Human Activity Recognition and Monitoring Using Smartphones", SpringerLink, 5th International Conference on Biomedical Engineering in Vietnam

[6] Sepp Hochreiter, Jurgen Schmidhuber, "LONG SHORT-TERM MEMORY", ResearchGate, Neural Computation in December 1997

[7] Klaus Greff, Rupesh K. Srivastava, Jan Koutn ´ık, Bas R. Steunebrink, J ¨urgen Schmidhuber, "LSTM: A Search Space Odyssey", arXiv:1503.04069

[8] M. H. M. Noor, Z. Salcic, K. I-K. Wang, "Dynamic sliding window method for physical activity recognition using a single tri-axial accelerometer", 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA)

[9] Javier Ortiz Laguna, Angel García Olaya, Daniel Borrajo, "A Dynamic Sliding Window Approach for Activity Recognition", UMAP 2011: User Modeling, Adaption and Personalization

[10] Ralf C. Staudemeyer, Eric Rothstein Morris, "Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks", arXiv:1909.09586

[11] Zachary C. Lipton, John Berkowitz, Charles Elkan, "A Critical Review of Recurrent Neural Networks for Sequence Learning", arXiv:1506.00019

[12] Maximilian Du, "Improving LSTM Neural Networks for Better Short-Term Wind Power Predictions", arXiv:1907.00489

[13] Alex Graves, "Generating Sequences With Recurrent Neural Networks", arXiv:1308.0850

[14] Irfan Ahmad Khan, Adnan Akber, Yinliang Xu, "Sliding Window Regression based Short-Term Load Forecasting of a Multi-Area Power System", arXiv:1905.08111

[15] Peter Faymonville, Bernd Finkbeiner, Maximilian Schwenger, Hazem Torfah, "Real-time Stream-based Monitoring", arXiv:1711.03829

[16] Antonio Bevilacqua, Kyle MacDonald, Aamina Rangarej, Venessa Widjaya, Brian Caulfield, Tahar Kechadi, "Human Activity Recognition with Convolutional Neural Networks", arXiv:1906.01935

[17] Daniil Osokin, "Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose", arXiv:1811.12004

[18] Deepika Singh, Erinc Merdiven, Ismini Psychoula, Johannes Kropf, Sten Hanke, Matthieu Geist, Andreas Holzinger, "Human Activity Recognition using Recurrent Neural Networks", arXiv:1804.07144

[19] Schalk Wilhelm Pienaar, Reza Malekian, "Human Activity Recognition Using LSTM-RNN Deep Neural Network Architecture", arXiv:1905.00599

[20] Oscar D. Lara, Miguel A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors", IEEE Communications Surveys & Tutorials

[21] Alex Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network", arXiv:1808.03314

[22] 'OpenCV', https://opencv.org/.