Human action and activity recognition using deep neural networks

*https://www.researchgate.net/publication/347443280

Uday Kumar Dussa
Department of Computing
Bournemouth University
Bournemouth, United Kingdom
s5229871@bournemouth.ac.uk

***Abstract*—Human activity recognition has turned into an exciting area in domains, such as health care, that resulted in using multiple approaches to solve a problem accurately and in the most efficient way possible. For the most prolonged period, standard means used to extract hand-crafted features from raw data signals. The classification was done using support vector machines and models related to hidden Markov. To improve the extraction of these features in a fashionable and highly automated manner, the concept of deep learning is integrated. The Deep-learning model comes in different modes, for example, in the structure or model of long-short term memory that can capture a data's sequential nature, thus outperforming other existing techniques. However, these models have their drawbacks because they often require significantly longer training and suffer from distant memory loss. With the emergence of convolutional neural networks, these drawbacks can overcome. The convolutional neural networks take the least time to get trained compared to long, short-term memory. They have a remarkable ability to capture dependencies that are long term compared to long short-term memory. The pairing of these two models with a convolutional autoencoder often removes the effects of noise and reduces a problem's complexity. When their robustness is tested regarding temporary data, convolutional neural networks have always outperformed the long-short-term memory. These networks can be used in building smart environments to detect any falls. This paper also sheds light on the various pros and cons that result in applying or using deep neural interconnections in human activity and movement recognition.**

***Keywords*—Human activity recognition; Support Vector Machines(SVM); Convolutional Neural Networks(CNN); Deep-learning model; Long Short-Term Memory(LSTM); Convolutional Auto-Encoder.**

I. Introduction

The analysis related to human activity turned out to be one the most crucial challenging issues that have attracted the computer vision community for quite some time in the current years. This analysis has resulted in many applications amounting from activity comprehension for intelligent surveillance systems to human-computer interaction improvement. The various approaches in use have significantly and effectively demonstrated performance in recognizing individuals' actions [1]. To acknowledge these interactions and group activities, the researcher would require information that is way more than just motion related to individuals. This identifies the research as a challenging one due to various individual aspects such as the intra-class variation of human-related activities, which are credited to the distinctions in visual appearance, fluctuations of subject movement, and changes associated with a viewpoint. To curb these challenges, the approaches on human activity recognition have chosen to place their focus on data about context, which can be recognized as information that is indirectly identified with the human action in itself. However, these actions can improve the general target focused action acknowledgment [2]. The existent approaches for human activity recognition will utilize people as the context without having any more extravagant context data that may incorporate sensitive information, which is where the activity took place and the whereabouts of the person within the scene. The previous approaches and methods have used context directly as input features to classifiers and support vector machines. This context has been incorporated through the aid of probabilistic models [3]. Therefore, proving the little effort employed in deep-learning models and convolution neural networks could capture human activity recognition. The deep-learning models tend to have the ability to incorporate various systems and multiple context sources to their deep multi-level structure [2]. They also have a more excellent capability of probabilistic reasoning alongside their ability to integrate hidden units for the sole purpose of synthesizing high-level representations of input features that could potentially be raw. This paper proposes a deep neural network model that recognizes human activity by exploiting its capacity to reason probabilistically while incorporating multiple context information sources.

II. Literature review

In the past, human activity recognition was usually performed through the aid of spatial features and support vector machines to classify the representation of various elements that could either be dense or sparse points that are spatial on histograms. Later, the convolutional neural networks eventually gained momentum since it has shown relevant features that

the deep learning method was unable to deliver, outperforming this method. Human activity recognition has real-world applications [4]. Various low-level features are often introduced in human activity recognition systems as a means to describe activity observation. A component such as local space-time can or is often used to represent how humans move while observing in a video [5]. These representations are integrated with support vector machine classification schemes that are meant for recognition. On the other hand, they proposed a feature related to space- time and the Spatio-temporal quality as the primary descriptor for human-related motion. The extraction of features that often seem discriminative from raw data has proven to be a critical and challenging task for Human activity recognition. Its regularly utilized features frequently incorporate natural signal measurements that in- clude mean and variance of different time sequences and trans- formation coding symbolic representations. The classification methods, such as Decision trees, k-Nearest Neighbours, and Support Vector Machines, often undergo some training to dis- tinguish various activities, utilizing the handmade highlights. Different classification methods that work in combination with multiple learning algorithms such as logistic regression, Decision trees, and multilayer perceptron can be integrated to enhance recognition accuracy. Therefore, significantly improving activity contrasts with what every individual calculation can accomplish with its extensively shallow highlights. Deep Learning Approaches such as restricted Boltzmann machine, deep belief networks, and interconnected convolutional systems have been explored to create more generalized learning methods that could efficiently extract features directly from data entered. It can be achieved by introducing added nodes and added layers for grouping that increase computation complexity. The convolutional neural network model applies max-pooling layers on highlight location from a crude contribution to request to deliver scale-invariant highlights. It is then acquainted with a neuro-shrouded layer to merge features from multiple different channels where another Soft- Max layer is used in the classification result. Auto-Encoder Long-term Recruitment convolution network model used a deep learning approach into human recognition action [6]. The autoencoder is mainly used to represent learning and remove any available inherent noise. This model's raw CSI time window is transformed into a latent space of two hundred and fifty-six features. Other improved models often reduce the latent space to one of about twelve features or even below. A deep-learning-based selective channel recognition system that effectively incorporated quality evaluation and selection is introduced [7]. With this model's use, channels are selected using an amplitude over a particular threshold while the rest of the media are neglected on the assumption that they were uninformative. Channel hopping is also used where the deep learning model circularly hops into the various selected channels combining the adjacent channels into extended channels that consist of higher bandwidth. A filter with a low pass that contains an approximation of 100Hz frequency cut-off is used to noise data [2].

III. Related work

The related work regarding human action and activity recognition is discussed. The Convolutional Neural Networks and a specific type of Recurrent Neural Networks (RNN) called Long Short term memory are used to implement human action and activity recognition. Human activities result from the significant variability in movements that lead to activities such as walking, standing, sitting, jumping, and complex activities such as eating. It is difficult to detect some complex activities due to various gestures related to the activities. So, a dataset consisting of multiple activities that reflects the reality should be considered. Therefore, the WISDM [8] and MHEALTH [9] datasets are selected. CNN is comprised of convolutional layers for extracting features from the input data. There are three layers in CNN, such as input layer, hidden layer, and outer layer. The middle layer can be a convolution, pooling, or normalization layer based on the arrangement. CNN is used in detecting and locating the objects in an image and describes image content automatically. RNN models are developed to learn from sequence data, such as sequences of observations over time. Long short-term memory network (LSTM) of RNN is used to overcome the general limitations in training a stable RNN on sequence data. In, LSTM layers blocks of memory are connected recursively in a cell of memory. Each cell consists of gates to specify when to forget the previously hidden layers of the memory cell and further update the cells, thus allowing the network to use the time data.

IV. METHODOLOGY

The various approaches to human activity recognition were tackled through filters, which were used to smoothen the data and remove noise by discarding the component of its first principle followed by an extraction phase of features by integrating some random domain expert knowledge. The support vector machine incorporated to facilitate the learning of how the selected features can be mapped onto a dimensional space of n values can finally be used to predict any of the problematic samples to gain insight into these samples. To completely perceive the human activity, we use an in-depth neural organization approach in conjunction with the Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and a probabilistic inferencing model shown in the figure.

The first step is the pre-processing step that detects the people's sequences from the above model. A convolutional model is learned through the aid of an inception V3 network for the sole purpose of classifying them in action-pose. Individuals who get detected from the context region are fed into the recurrent neural network model for sequence learning. The sequential context feature is often trained through a SoftMax classifier. It produces an activity score that belongs to every group of elements or individuals in the context. From the group activity label, an input sequence is often fed directly towards the convolutional neural network, the inceptionV3 network then the SoftMax classifier. A maximal posterior probability

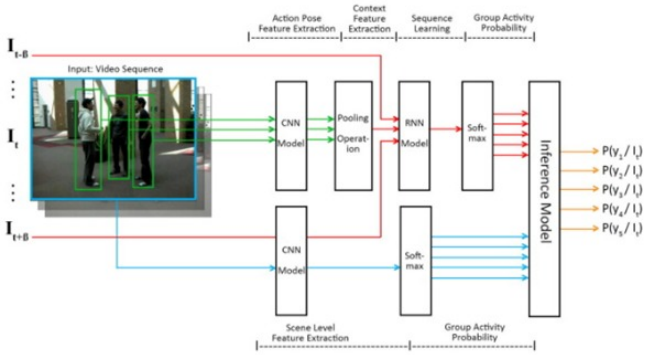

Fig. 1. Model

is used to classify the various group activities available to the probabilistic inference model.

V. Pros

Deep learning has benefits in its usage as far as human activity recognition is concerned. It is highly recommendable for the use of an application. The convolutional neural network has great potential in identifying essential features regarding the signals of human activity recognition. The local characteristics of the signs that represent the possibility of movements in human activity are frequently obtained by the lower layers of the processing unit. The higher layer regularly represents the information in a more theoretical way to describe or instead categorize the importance of various combinations of motion as each layer could normally constitute multiple convolutions and pooling operations that are often specified differently using different parameters. In comparison with artificial extraction, convolution neural networks play a crucial role in taking the various features in data and consider other aspects that eventually become more comprehensive. Therefore, deep learning brings a totally no need for feature engineering. Deep learning offers the best results relating to unstructured data. The data that belongs to organizations is often in unstructured formats since most of it usually exists in forms such as pictures, texts, and pdf files, making the overall data challenging to analyze while utilizing AI calculations. The utilization of profound learning calculations comes in handy since they get trained using various data formats, enabling them to get more insights, ranking deep learning as a crucial tool. Deep learning facilitates easy data labelling. It is because of its ability to excel even without any guidelines, unlike other machine learning algorithms. A deep learning algorithm would quickly and efficiently detect any human body physical anomalies at very early stages that not even human doctors can notice. As human beings need ample resting time and refuel their energy, they also tend to make mistakes that could be rendered as careless ones resulting from them being tired or even hungry. Once a deep learning model is trained, it could have the capability to perform thousands of repetitive tasks. Within the shortest period possible, it would take an actual human. A deep learning brain's quality never diminishes, rendering it efficient enough to deliver high-quality results.

VI. Cons

When we cannot determine the root of a neural network's solution, these networks at the core of deep learning are often black boxes, making it extremely difficult and challenging to look into these networks and see how they work. Like the human brain, a neural network's reasoning is integrated with the behaviour that mimics multiple thousands of simulated neurons that are arranged in hundreds of intricate connection of layers, which together form a complex web that receives inputs emerging from one given level to the next until a valid and overall output is produced. The back propagation plays quite a role in tweaking neurons' calculations at individual levels. The networks learn how to make output that is desirable at a significantly faster rate. Even if neuro networks often produce good and hot results, they lack transparency in their thinking mode. Coming up with solutions often brings a challenge in predicting when failures occur. Over fitting is yet another major issue with neural networks. Over fitting refers to a situation or an algorithm that models the data to be trained quite too well or algorithms that over- train a given model. Over fitting situations tend to occur when algorithms learn both the details and noise in training data to extents that negatively impact the models' performance when applied in real-life situations. Modern neuro networks tend to have large data sets that bring about a lot of noise, which affects the efficient function of neuron networks.

VII. Conclusion

In conclusion, human action acknowledgment deep neural network from a video with inclusion of motion and context features have effectively proven to catch the immediate surrounding of the object interested in both at local and global extents. Both Convolutional neural networks, and LSTM networks are best for learning features from raw sensor data and for predicting associated motion/movement. From the above results, deep learning has many potentials. It needs to overcome its various minor challenges before it gets to the point of considering a one versatile and salient tool. Additionally, multiple interests and enthusiasms towards deep learning are rapidly growing as today's real-world applications of this technology have been embraced. The model's performance was outstanding as multiple and complex activities were recognized by a large variability of movements and diversity of body parts.

Acknowledgement

The author wishes to thank Dr Philips Davies and Dr Festus Adedoyin in the Department of Computing, Bournemouth University,UK for their continuous help and support.

REFERENCES

[1] Tran,K.N.,Kakadiaris,I.A.,andShah,S.K.(2012).Part-basedmotion descriptor image for human action recognition. Pattern Recognition, 45(7):2562–2572.

[2] Wang,X.andJi,Q.(2015).Videoeventrecognitionwithdeephierarchi- cal context model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4418–4427.

[3] Choi, W. and Savarese, S. (2012). A unified framework for multi-target tracking and collective activity recognition In Computer Vision–ECCV 2012, pages 215–230. Springer.

[4] K. Fu, Y. (2016). Human activity recognition and prediction. Cham: Springer.

[5] Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local svm approach. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, volume 3, pages 32–36 Vol.3.

[6] Mandic, D. P., Chambers, J. A. (2001). Recurrent neural networks for prediction: Learning algorithms, architectures and stability. New York: John Wiley.

[7] Balas, V. E., In Roy, S. S., In Sharma, D., In Samui, P. (2019). Handbook of deep learning applications.

[8] G. M. Weiss, "Wisdm smartphone and smartwatch activity and biometrics dataset," in UCI Machine Learning Repository, WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set, 2019.

[9] O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, and C. Villalonga, "mHealthDroid: A novel framework for agile development of mobile health applications," in Proc. 6th Int. Work-Conf. Ambient Assist. Living Act. Ageing (IWAAL), Belfast, U.K., Dec. 2014, pp. 91–98