Research on Human Gesture Recognition Algorithm Based on Multi-Scale Sparse Neural Network

Lexuan Huang[1], Shenggang Yan[2,*] and Jianguo Liu[2]
[1]*Shanghai United International School, Gubei Secondary Campus, Shanghai 201103, China*
[2]*School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China*

Keywords: Human Gesture Recognition, Multiscale Sparse Neural Networks, Time Series Data.

Abstract: Human posture classification, as a part of human activity recognition, has a wide range of applications in medical treatment, nursing, sports injury protection, safety monitoring, and other fields. Along with the increasing hardware computing power and storage performance, an extensive amount of large data can be stored and computed as time series in real-time. But in the face of vast amount of data generated by inertial sensors, there is still a lot of room for development in real-time data processing. Multiscale Sparse Neural Networks are in this paper, realize the judgment and classification of the posture of the human body. The data is the human body pose by the nine-axis inertial sensor and the huge amount of data and real-time calculation have been taken into account. Compared to other methods, ours performs state-of-the-art on the WISDM and UCI-HAR datasets, and achieves good results on our own datasets.

1 INTRODUCTION

With the rapid development of electronic and information technology, mobile phones have become indispensable personal products. The sensors in mobile phones, such as built-in gyroscopes and magnetometers, can help people do assorted online activities. These sensors integrate the tools that people need into a small and convenient mobile phone. In particular, the acceleration-based step counter can help individuals understand their physical activities, to achieve multiple purposes such as calorie calculation (Nweke et al., 2018; Dirac et al., 2013). Although the existing step counter is more accurate for calculating the number of steps, it will mistakenly judge other non-walking situations as the walking state due to the change of the relative position. Based on this question, the following thoughts can be drawn. Can the sensor be used to distinguish the user's action status? So as to distinguish different travel modes, and at the same time, it can also increase the judgment accuracy of mobile phones and other portable devices in various situations. Higher requirements are put forward on low computational complexity and high real-time performance. In practice, the measurement of human body posture through a computer is not only applied to pedometers but also widely used in the following scenarios.

1.1 Application of Human Body Posture Classification

With the development of computer intelligence, humans' understanding of machines continues to make breakthroughs, and the development of more advanced human-like machines complements accurate researches on human cognition. The human body posture classification is an important part of human behavior cognition. The human body posture estimation refers to the accurate detection and prediction of an individual's posture (Ma et al., 2017). It has a wide range of application scenarios and a broad development market.

1.2 Time Series Data

In human body posture classification, the input samples and tests are both time series data. Time series data is a type of data collected at different times and serves to describe changes in the phenomenon over time. This type of data reflects the change in the state or degree of a certain object or phenomenon over time. With time characteristics, it can intuitively reflect the transformations of data over time. After storing the human body posture data as a time series, it is easier to observe the human body posture data that alters due to time changes in motion (Xia et al., 2012). This kind of data is widely present in the

Huang, L., Yan, S. and Liu, J.
Research on Human Gesture Recognition Algorithm Based on Multi-Scale Sparse Neural Network.
DOI: 10.5220/0011951500003612
In *Proceedings of the 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 457-464
ISBN: 978-989-758-622-4; ISSN: 2975-9463

sensor sequence stored as time goes on. For example, the observation and maintenance of bridges: In recent years, extreme weather and floods have become more frequent, and the damage caused has greatly intensified (Nishani and Çiço, 2017). By installing sensors in bridges, researchers can use the collected time series and the constructed model to analyze changes in the state of the infrastructure, effective early maintenance, and warning measures (Omenzetter and Brownjohn, 2006). This kind of data will also be utilized in econometric models, such as a country's GDP data. Through time series data, researchers and experts can understand the trend of GDP growth over the years.

Time series data can be divided into stationary processes, de-trend stationary processes, and differential stationary processes. For example, in the steel wire manufactured by the twisting method, the random process in which the diameter of the steel wire does not change with the lapse of time is stable; when the water droplets penetrate the stone, the water droplets continuously invade the stones, and the amount of stone reduction has an upward trend. The statistical characteristics of time series detrending can be obtained; annual rainfall characteristics pass trend and seasonality, and stable rainfall characteristics data can be obtained after differential conversion, equating, a data set with stable mean and variance, which is a differential stationary process.

In order to solve these problems, Multi-scale and sparse neural networks are studied in this paper, different from traditional algorithms for human pose detection. Our proposed method has good adaptability at large data scale by improving the sparse detection ability of the network. The difference from the existing algorithms (Golyandina, 2020; Perraudin et al., 2017; Korenberg and Paarmann, 1991) is that the receptive field is self-adapted in the configuration of our algorithm. The combination of multi-scale and sparseness on the network brings a new dimension of representation at the level of real-time data. It shows good characteristics when the data is collected by the nine-axis sensor mounted on the human body.

1.3 Challenges

At present, most human posture monitoring devices are based on the information of video images. This method can recognize the human joint structure through images and construct 2D or 3D bones (do Rosário, 2014; Le and Nguyen, 2013). It has been well applied in some fields. Based on the research purpose of judging and classifying human posture, this paper selects high-precision sensors to complete data acquisition. For example, a sports Bracelet uses a gait cycle estimation algorithm (Moe-Nilssen and Helbostad, 2014). However, for the problem to be solved in this paper, the traditional algorithm will have the problem of false recognition, less recognition of human motion state, and cannot make effective judgments on bus travel and car travel because it cannot distinguish the motion state. Once when I checked my mobile phone by bus, I found that the number of steps on the counter was increasing, which was caused by the sensor misjudging the bumpiness of the bus as walking. Secondly, the existing and widely used gait cycle estimation algorithms not only cannot achieve multi-objective classification and judge a variety of travel patterns but also cannot process a large number of data generated in our research process.

At the same time, there will be some challenges when analysing time series data. When the collected time series data is incomplete, the trend about time obtained by analysing this incomplete data is very high, which may be wrong or biased. For example, collecting the water level change of a river under the influence of the tide, but only collecting the data in the dry season, or the imbalance of various state data will affect the data classification results. In the process of this study, three kinds of sensor data, namely three-axis accelerometer, gyroscope, and three-axis angular velocity sensor, are used for calculation. The amount of data is large and the characteristics are complex. The data collected and analyzed by the traditional algorithm cannot meet the requirements of this study.

2 RELATED WORK

The continuous development of deep learning leads to numerous developments and achievements in human posture classification. At the very beginning, machine learning played an important role. The Support Vector Machine (SVM) (Byvatov and Schneider, 2003) is one of the most widely used machine learning algorithms. SVM analyses the data-through a linear decision hyperplane. During training, the linear decision hyperplane is trained and adjusted in order to separate data with different labels (Chathuramali and Rodrigo, 2012; Tharwat et al., 2018). In the article (Chathuramali and Rodrigo, 2012), the author used images after feature extraction as the input of SVM. As a result, the SVM is quite computationally cost-effective and accurate in high-dimensional vector space. K-Nearest Neighbors (K-

NN) is another commonly used classifier. K-NN only has one parameter k for variant how many nearby samples will be based on for classifying the type of the test sample, and the choice of K will have a significant effect on the especially high dimensional data processing (Akilandasowmya et al., 2015). In the article (Li et al., 2012), the author applied K-NN to human posture classification and result in high accuracy. For the KTH dataset, the average accuracy reaches 91.4%. Thereafter, deep learning algorithms, as a type of more complex machine learning, are also used in human posture classification. Back Propagation Neuron Network (BP neuron network) algorithm is a multi-layer feedforward network, training with the method based on error back propagation (Jin et al., 2000). Researchers use the algorithm discovered by Paul (Jain and Kanhangad, 2017) to learn the features, replacing the manual feature extraction. LeCun et al. applies the BP neuron network to handwritten digit recognition, displaying the possibility that BP Neuron Network can be used in Image recognition without a complicated pre-processing stage. Due to BP Neuron Network updates the weighted based on the direction with the most descent gradient of error, it is inevitable that the error will finalize in a minimum state in a particular section and cannot reduce to a degree that has the smallest error (Guo et al., 2020). This is the most significant disadvantage of the BP Neuron Network. Along with promoting computational ability, CNN is developed while other algorithms that combine with CNN, such as LSTM-CNN, are also created. Before introducing the LSTM, a discussion about RNN has to be made. The recurrent Neural Network (RNN) is a neuron network that is used to manage the series-changing data. This kind of neural network can be generally utilized in language modeling, speech recognition, machine translation, and other applications. The Long Shot-Term Memory (LSTM) is typically designed to solve the problem of long-term dependency in RNN. LSTM-CNN then combines the advantages of both CNN and LSTM neuron networks. This neuron network can be flexibly used in visual signal processing that involves continuous input and output, applying CNN for input feature extraction and LSTM for the series prediction. The technology developed is employed in activity recognition, image and video description, natural language processing, and so on. LSTM-CNN maintains a high accuracy as the result. According to the article [21], the author applies LSTM-CNN to human posture recognition and acquires the percentage accuracy of 95.78%, 95.85%, and 92.63% in the public dataset UCI-HAR, WISDM, and OPPORTUNITY correspondingly, proving the superiority of this type of neuron network.

3 MATERIALS AND METHODS

3.1 Data Description

To train the neuron network with human posture data, this paper uses two public datasets and a self-collected dataset, which both will be introduced in the following passage. For the public dataset, WISDM [22] and UCI-HAR [23] are selected, and both of them include six different human postures. Comparing to WISDM, UCI-HAR has more data-collecting volunteers, represents the vast majority. Although the WISDM dataset has more data collected. The self-collected data set is based on the nine-axis sensor, collecting the human posture data on different transportation.

WISDM: This is an integrated and ready-to-be-used dataset, including over one million data collected from 36 different users. All of the data is collected through the sensor from an Android smartphone, which is required to be placed in a pants pocket while doing the six following activity——walking, jogging, going upstairs, and going downstairs, sitting and standing. When volunteers do these activities, the sensor collects the data in the frequency of 20Hz and records them in time series. The distribution of data is shown in Table 1.

Table 1: WISDM: Wireless Sensor Data Mining.

Class Distribution	Number of example	percentage
Walking	424,000	38.6%
Jogging	342,177	31.2%
Upstairs	122,869	11.2%
Downstairs	100,427	9.1%
Sitting	59,939	5.5%
Standing	48,395	4.4%

UCI-HAR: This dataset is collected from thirty volunteers aged between 19~48 years old, and each of them placed the phone at their waist while carrying out six daily activities similar to WISDM, namely walking, going upstairs, going downstairs, sitting, standing, and lying down. The distribution is shown in Table 2. With the sensor inside the phone, the three-axis acceleration and three-axis angular speed are collected at a frequency of 50Hz. After data collection, the researchers denoise the data and label them accordingly to the video recorded during the collection process, then separate them into two

groups. 70% of the data goes into the training set and the rest 30% is in the testing set.

Table 2: UCI-HAR: Human Activity Recognition database

Class Distribution	Number of example	percentage
Walking	122,091	16.3%
Walking-upstairs	116,707	15.6%
Walking-downstairs	107,961	14.4%
Sitting	138,105	16.9%
Standing	136,865	18.5%
Laying	136,865	18.3%

Self-collected dataset: While collecting the data, Raspberry Pi with a nine-axis sensor is used. Comparing to the sensor in the smartphone, Raspberry Pi (Figure 1) can carry out a wider range of sensors (Figure 2) that can be customized, resulting in the availability of more types of data. The Raspberry Pi was placed in the pants pocket, where the smartphones are usually placed, to collect the data that has similar features to the one which the phone will collect. The types of data collected are x, y, z, three-axis acceleration, angular speed, and gyro value, a total of nine types of data, collecting at the frequency of 20Hz. During the data collection, six modes of traveling are chosen, which are walking, running, cycling, bus, metro, and car. Each activity lasts for around one hour to collect complete sets of data. The distribution of the self-collected dataset is quite even, listing in Table 3.

Table 3: Self-collected dataset.

Class Distribution	Number of example	percentage
Walking	89349	17.4%
Running	74098	14.5%
Cycling	90187	17.6%
By bus	60982	11.9%
By subway	87623	17.1%
By car	108761	21.3%

3.2 Data Preprocessing

In the data collected, there will usually be some anomalies and missing values, which will affect the training of the model. To get the best result of network training, data need to go through pre-processing ensuring that the anomalies are removed and the missing values are replenished, which is especially important for the self-collected data. This paper did the following data pre-processing:

Data Normalization: Normalization is the process of reaching the minimal redundancy that is integral to structure the data. Generally speaking, the dataset will be divided into several groups, and the relationship between groups will be defined. In that case, modification and deletion of one single group will be spread to others with a defined relationship. When a relational database has a higher normal form, it will have a higher resistance to anomalies and be convenient to process.

Missing Value Processing: Since the time series data is used in various fields, it is inevitable that some of the data will be lost during the transmission. These missing data will affect the overall effectiveness of network training. There are two mainstream missing value processing methods. The first one is refilling data based on the statistic pattern. A researcher refill by last, mode, or mean. The second one is deleting the group of samples that miss the value. In the self-collected dataset, the missing value is caused by the deletion of anomalies. Filling in the missing value is one of the works done in this paper through the mean of the previous and the next data.

Figure 1: Raspberry Pie Chip.

Figure 2: Nine-axis Sensor.

3.3 Network Structure

A neural network architecture (Figure 3) is designed in this paper and, in particular, is added to a simple and efficient residual processing pipeline. Different from the existing methods to improve the feature capture accuracy, the method we propose can capture

the dependency information well in a limited level. Compared with [24], [25], [26], the defect is solved in handling timing problem in CNN.

Residual network is widely used in deep neural network, it can solve the problem of gradient disappearance and gradient explosion caused by too many layers (Guo et al., 2020). Expanding the receptive field of a neural node generally increases progressively using the inflation rate. Expanding the receptive field of neural nodes generally increases the use of the expansion rate progressively, the network depth will increase accordingly, and the stability of the network will be enhanced due to the introduction of residual connections. Figure 4 shows the residual structure used in this paper, including the output of $\boldsymbol{F}$ and the identity map $\boldsymbol{X}$ in its final output $\mathbf{O}$.

$$o = \sigma(F(x) + x) \quad (1)$$

A human pose classification model is established for long-term sparse sequence signals. In this paper, the three-level residual network structure is stacked to increase the network receptive field, and the structure is more stable in this network. The sequence dimension is 800*1, which is the input dimension of this network input. After sufficient experiments, the filter uses 5*1 and the sliding step size in the convolutional layer is 1. The filter is set to 32 and 64 in the pre- and post-stage filters. It can be seen that the larger receptive field greatly reduces the computational cost compared to the traditional CNN network in a small number of base-level networks through the stacked network. At the same time, it has good feature information capture ability in longer sensor time series.

4 EXPERIMENT RESULT

We implemented our model with the TensorFlow framework. Make a comparison with different algorithms on different datasets. Validate the effectiveness of the network by validating experiments, evaluating and analyzing the results on our own datasets. Finally, the confusion matrix is obtained on the training set, validation set, test set and the total classification results on different datasets.

Differences In terms of different data sets, the public data sets are all collected from the real environment, but the obvious data differences are reflected in the sample types and action characteristics. The parts that are easily confused between different actions are more obvious in the sparse features in the own data set. There is a clear difference in the sparse feature representation using the residual network structure in Figure 4, and the results of processing these data are shown in Figures 5, 6, and 7. The experimental results show that the algorithm in this paper shows good performance in the public data sets WISDM and UCI-HAR, and the performance is excellent in the own datasets.

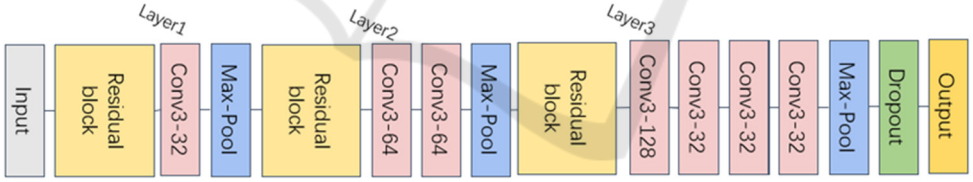

Figure 3: network architecture.

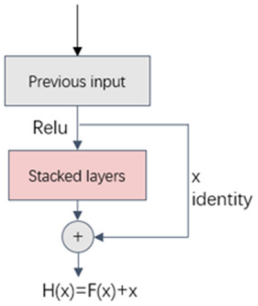

Figure 4: Details of the residual block.

Comparison between different algorithms, LSTM-CNN, CNN, SVM, J48, and multilayer perceptron were used to compute the same data. Accuracy and training time are the detection criteria, as shown in Table 4. It can be seen that the accuracy rate has been improved using the method in this paper compared with other algorithms. Exhibits larger gaps using these algorithms on own datasets due to larger data size and smaller class gaps.

Figure 5: The confusion matrix uses the WISDM dataset on the algorithm in this paper(The abscissa is the target classification and the ordinate is the output classification)

Figure 6: The confusion matrix uses the UCI-HAR dataset on the algorithm in this paper(The abscissa is the target classification and the ordinate is the output classification)

Figure 7: The confusion matrix uses the Self-collected dataset on the algorithm in this paper(The abscissa is the target classification and the ordinate is the output classification)

Table 4: Comparison of accuracy recognition and training time between this algorithm and other algorithms

datasets / Model	Recognition Accuracy /%			Training time/s		
	WISDM	UCI-HAR	Self-collected	WISDM	UCI-HAR	Self-collected
LSTM-CNN	90.2	92.4	86.5	14.253	15.3635	19.461
J48	81.1	85.1	79.6	10.263	9.683	11.254
multilayer perceptron	88.3	91.7	82.6	11.322	10.256	12.364
SVM	87.6	91.9	84.5	9.651	7.229	10.251
CNN	90.6	91.2	85.7	20.356	13.529	16.321
The algorithm of this paper	96.3	93.33	97.42	7.586	6.241	8.254

5 CONCLUSION

To solve the above problems, as well as multi-object classification, multi-dimensional data processing, and accurate half-segment results, the following work is done.

(1)All kinds of data are collected and preprocessed. In the collection process, the collection time of each sample is controlled in an interval greater than one hour to ensure the integrity of the data, and to reduce the proportion of the overall data when abnormal data occurs, thereby reducing its impact on the data set. In the preprocessing, the data is standardized and missing values are processed.

(2)A deep multi-scale neural network is proposed as our algorithm, and the fast and accurate extraction of data features in low signal-to-noise ratio data is the advantage of this algorithm. Find a balance between high resolution and high real-time through different receptive fields. Reasonable use of pooling and up-sampling techniques can improve performance while reducing the amount of computation. The receptive field changes continuously during the calculation process in the blocks between layers, so as to achieve the purpose of capturing local features in a large amount of data.

(3) Use the public data set as input to test this model, because the public data set has the characteristics of large quantity, standard collection, and is often used, so it has a high degree of reference when testing the model. In addition, the model has achieved high accuracy in multiple public data sets, which proves that the model has sufficient versatility.

(4) Comparing the results obtained by running multiple public data sets of this model with the results obtained by various existing models, a higher accuracy has been achieved, which proves the practical application value of this model.

Based on real-life problems, this research built a sensor collection device to process relevant data. Carrying out network model research and experiments based on CNN and LSTM-CNN networks and other, and verify network models based on WISDM and UCI-HAR data sets, and elaborate and demonstrate in the papers respectively to realize human posture detection and classification. This produces the intelligent recognition of different transportation vehicles when going home in real life. In this process, a large number of documents were investigated, the world-class research methods were summarized, and scientific research capabilities were exercised.

REFERENCES

Nweke, Henry Friday, et al. 2018. Deep Learning Algorithms for Human Activity Recognition Using Mobile and Wearable Sensor Networks: State of the Art and Research Challenges. *Expert Systems with Applications*, pages. 233–261.

Dirac G, Leone A, Siciliano P. 2013. Human posture recognition with a time-of-flight 3D sensor for in-home applications. *Expert Systems with Applications*, 40(2): 744-751.

Ma C, Li W, Gravina R, et al. 2017. Posture detection based on smart cushion for wheelchair users. *Sensors*, 17(4): 719.

Xia L, Chen C C, Aggarwal J K. 2012. View invariant human action recognition using histograms of 3d joints. *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. pages 20-27.

Nishani E, Çiço B. 2017. Computer vision approaches based on deep learning and neural networks: Deep neural networks for video analysis of human pose estimation. *2017 6th Mediterranean Conference on Embedded Computing (MECO)*. page 1-4.

Omenzetter P, Brownjohn J M W. 2006. Application of time series analysis for bridge monitoring. *Smart Materials and Structures*, 15(1): 129.

Golyandina N. 2020. Particularities and commonalities of singular spectrum analysis as a method of time series

analysis and signal processing. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(4): e1487.

Grassi F, Loukas A, Perraudin N, et al. 2017. A time-vertex signal processing framework: Scalable processing and meaningful representations for time-series on graphs. *IEEE Transactions on Signal Processing*, 66(3): 817-829.

Korenberg M J, Paarmann L D. 1991. Orthogonal approaches to time-series analysis and system identification. *IEEE Signal Processing Magazine*, 8(3): 29-43.

do Rosário J L P. 2014. Photographic analysis of human posture: a literature review. *Journal of bodywork and movement therapies*, 18(1): 56-61.

Le T L, Nguyen M Q. 2013. Human posture recognition using human skeleton provided by Kinect. *2013 international conference on computing, management and telecommunications*. pages: 340-345.

Moe-Nilssen R, Helbostad J L. 2014. Estimation of gait cycle characteristics by trunk accelerometry. *Journal of biomechanics*, 37(1): 121-126.

Byvatov E, Schneider G. 2003. Support vector machine applications in bioinformatics. *Applied bioinformatics*, 2(2): 67-77.

Chathuramali K G M, Rodrigo R. 2012. Faster human activity recognition with SVM. *International conference on advances in ICT for emerging regions* pages: 197-203.

Tharwat A, Mahdi H, Elhoseny M, et al. 2018. Recognizing human activity in mobile crowdsensing environment using optimized k-NN algorithm. *Expert Systems with Applications*, 107: 32-44.

Akilandasowmya G, Sathiya P, AnandhaKumar P. 2015. Human action analysis using K-NN classifier. *2015 Seventh international conference on advanced computing*. pages: 1-7.

Li J, Cheng J, Shi J, et al. 2012. Brief introduction of back propagation (BP) neural network algorithm and its improvement. *Advances in computer science and information engineering*. Heidelberg, pages: 553-558.

Jin W, Li Z J, Wei L S, et al. 2000. The improvements of BP neural network learning algorithm. *WCC 2000-ICSP 2000. 2000 5th international conference on signal processing proceedings*. 3: 1647-1649.

Jain A, Kanhangad V. 2017. Human activity classification in smartphones using accelerometer and gyroscope sensors. *IEEE Sensors Journal*, 18(3): 1169-1177.

Guo C, Fan B, Zhang Q, et al. 2020. Augfpn: Improving multi-scale feature learning for object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pages: 12595-12604.