Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/2023305

Sensor-based gesture recognition with convolutional neural networks

Ran Bi

School of Glasgow, University of Electronical Science and Technology of China, Chengdu, China

2510938b@student.gla.ac.uk

Abstract. Sensor-based gesture recognition is an active research field of great significance with a wide range of applications in control systems for virtual reality, medical monitoring, and abnormal behavior determination. This problem has drawn the attention from both the academia and industry and many methods are proposed in the literature. Recently, deep learning has been widely applied in sensor-based gesture recognition and achieved good effects. In this paper, we proposed a classifier model based on Convolutional Neural Network (CNN) and applied it to EMG-based and smartphone-based datasets, respectively. For these two datasets, our model both achieved better classification accuracies than traditional machine learning models, with the results of approximately 97% and 72% accuracies, respectively. We also analyzed the effects of different parameters on the results of the proposed CNN model.

Keywords: sensor data, gesture recognition, convolutional neural network

1. Introduction

With the increasing demands of convenient communication, sensor-based gesture recognition has been one of the foremost research directions in recent years. Communicating through gestures can not only happen between humans, but also can be used in human-computer interaction. For us humans, gestures are more expressive and impressive than other interaction approaches, without the usage of languages. At the same time, gestures can be quickly analyzed and understood by computers. Due to its variable applicability in different domains, gesture recognition is widely used in monitoring system [1], health care [2], intelligent living [3], and so forth. These gesture recognition applications can be categorized into touch-based and touchless methods. Touch-based methods always collect information or signal from a device attached to the body. In contrast, touchless recognition is mainly based on the visual analysis of human movements.

Existing gesture recognition methods can be classified into based on surface electromyography (sEMG) and smartphone, based on the data set collection methods. On the other hand, they can be divided into traditional and deep learning methods, based on recognition models. sEMG is the combined effect of the superficial muscle and electrical activity of nerve trunk on the surface of the skin which can be caught by the surface electrode placed on the skin. It is simple, non-invasive and safe. However, the number of hand gestures that can be recognized is quite limited and the overall accuracy cannot satisfy the requirements. Touchscreen-enabled smart phones with motion sensors are also able to collect

DOI: 10.54254/2755-2721/4/2023305

information through the human movements. Compared with direct electrical signals, its recognition is much more complicated. However, it obtains multiple features and achieve a better accuracy without the attached equipment. Some approaches use traditional machine learning algorithms, which show an excellent performance with preprocessed data. But when the data set becomes much bigger or contains more features, the models contained in machine learning could not work well. As a result, some strategies based on deep learning are proposed to improve the recognition accuracy, including 3-dimensional deep convolutional network (C3D) [4], recurrent neural network [5], and a transfer learning strategy [6]. Deep learning decreases the difficulty of handling a large number of data and requires less professional design of function. Moreover, its accuracy achieves a higher level when applying adequate data and appropriate algorithms. Although these methods show better consequences than traditional models, the average accuracy is still lower than expected.

In light of the existing issues, we proposed a gesture recognition model based on Convolutional Neural Network (CNN) that had been applied to a EMG-based dataset and a smartphone-based dataset in this study. The EMG-based dataset is referred as dataset 1 and the smartphone-based dataset is referred as dataset 2 in the following content. This model integrates the simplicity and operability of deep learning, achieving better classification accuracies than traditional machine learning models. With suitable models, the accuracy for these two data sets can reach 98% and 72%, respectively, which is higher than 97% and 57% of machine learning baselines. The reason for the huge difference in smartphone-based data is that CNN presents better performance on more complex data. Apart from that, we also analyze the effect of different hyper-parameters on the effect of convolutional neural network models on dataset 2. We concluded that the number of 2D convolutional layers has the most influence on final results, while changing sampling points, k-fold, dropout probability barely influences the outcome.

The remainder of the paper is organized as follows. Section 2 reviews the related work such as machine learning and deep learning algorithm. Section 3 gives the data description. Section 4 described the detailed model and experiment description. Section 6 gives a conclusion.

2. Related work

Over the last decade, gesture recognition has attracted a wide attention due to the rapid development and widespread application of sensor technologies and its universal application [7]. Gesture recognition can automatically classify gestures with different data formats, e.g., EMG-based and smartphone-based data. The recognition process can be realized by machine learning or deep learning, of which the accuracy differs a lot. First, confirm that you have the correct template for your paper size.

2.1. EMG-based gesture recognition

2.1.1. Traditional methods. When using the traditional machine learning algorithm, Rotation Forest (RoF) and Extreme Learning Machine (ELM) were integrated to enhance the generalization stability and accuracy in [8]. At first, the active movement segments were pre-processed by sliding window and 104 features were extracted from each segment. The space dimension of the most contributed results would be reduced by Support Vector Machines-Recursive Feature Elimination (SVM-RFE) model. At last, the integrated RoF-ELM classifier was built and tested on sEMG signals. Compared with the single RoF or ELM model, the accuracy of this method is the highest that can reach approximately 91.11% with a relatively short running time. Through the process of windowing, capturing more features brings more redundancy and unrelated information that is adverse to gesture recognition. One of the Machine Learning model, Support Vector Machine (SVM), was also used to classify 13 hand gestures [4]. Six features in the time domain, frequency domain and time-frequency domain are captured and the average accuracy of SVM exceeds 95%.

2.1.2. Deep learning methods. On the basis of machine learning methods, new methods have emerged more recently, especially those deep learning models [9][10][11]. Some are the combination of

DOI: 10.54254/2755-2721/4/2023305

traditional methods and deep learning; others are completely new strategies based on deep learning models [12][13].

In [6], a Transfer Learning (TL) strategy based on convolutional neural network was proposed to improve the performance of instantaneous gesture recognition. This method expanded the surface electromyography (sEMG) application domain to new subjects and new hand gestures. The accuracy of the proposed strategy was improved by 18.7% and 8.74%, respectively while using up to three repetitive gestures. Furthermore, it validated the portability of spatial characteristics and the effectiveness of this strategy for new gesture recognition.

A framework for recognizing multiple hand gestures in two stages was proposed in [14]. First, the hand gestures were split into multiple superclasses by y conventional time-domain features. Then, Multivariate Variational Mode Decomposition (MVMD) was applied to sEMG signals to extract the temporal and spatial characteristics of multiple channel signals. At last, CNN was used to train the input signal. The average accuracy for this framework could be approximately 90% while choosing 52 hand gestures as input signals. Multi-Layer Perceptron (MLP) and multi-channel Convolutional Neural Network (multi-channel CNN) were applied to classify 13 hand gestures from six muscles of forearm and hand. There was no distinct difference between MLP and multi-channel CNN with above 95% accuracy. The loss function of multi-channel CNN approached zero rapidly both in the train period and test that indicates the good performance of the model. CNN was also applied to improve the performance of hand movements prediction, with segmented sEMG signals [15]. In that study, the spectrum images of sEMG signals were obtained by Short-Time Fourier Transform (STFT). After that, these colorful spectrum images were trained with 50-layers Residual Networks (ResNet) that is a subclass of CNN. The training accuracy was 100% and test accuracy was 99.59%. In [16], CNN was implemented to analyze the influence of hyper-parameters from the sEMG signals collected on 18 subjects. It is shown that the learning rate had negligible effect on the result. However, some specific motions presented a better performance for all cases.

2.2. Smartphone-based gesture recognition

2.2.1. Traditional methods. A machine learning system was proposed to identify hand gestures based on the data collected from the front camera of a smartphone [17]. The machine learning algorithm implemented with Hu image moments represented an affordable compromise between precision and computing cost on mobile devices. The experimental results showed that it could correctly recognize 91.96% gestures for a whole of 4800 tests.

2.2.2. Deep learning methods. In [4], a new framework combined with 3D-CNN and Long-Short-Term-Memory (LSTM) were constructed to recognize human actions. The distinct information was integrated to generate a motion map using an iterative method. Then, an effective fusion scheme was introduced to blend spatial and temporal features. The average accuracy based on three well-known human action datasets were nearly 90%, which outperformed traditional models by about 4.5%. Another deep learning method called Bi-directional Long-Short Term Memory (BiLSTM) was used in [18], which combined with Skip-Chain Conditional Random Field (SCCRF). This hybrid approach can be divided into two phases. The first phase was to recognize the simultaneous motions using the BiLSTM technique, while in phase two, the interlaced activity was used to identify the interlaced activity. The average accuracy of the two-phase system using the smart home environment datasets could exceed 93%.

The Recurrent Neural Network (RNN) is a kind of artificial neural network with one module called Gated Recurrent Unit (GRU). Using convolutional and GRU layers, a deepGesture algorithm was proposed to recognize human activity [19]. Four convolutional layers firstly learned features automatically in raw data collected from the smart band device. Then these learned features were utilized as inputs of the GRU layer to capture the long-term dependency from the sequential data. The average precision of recognizing nine arm gestures could be approximately 95%.

DOI: 10.54254/2755-2721/4/2023305

3. Data description

The two datasets used in this experiment was related to gesture recognition with different data type. While dataset-1 included electrical signals, dataset-2 made use of time series data.

Dataset-1 was collected from muscle electrical signals placed on the arm. Eight sensors were placed on the skin of the arm, and each sensor would get eight electrical signal data. The structure of the sensor and its working schematic are shown in figure 1 and figure 2 respectively. As a result, a total of 64 data were included in one gesture. The 65th data of the training set is the gesture category. There are four kinds of gestures numbered 0, 1, 2, 3, representing rock, scissors, cloth, and OK. The corresponding gestures for four numbers are displayed in figure 3.

Figure 1. Mvo armband.

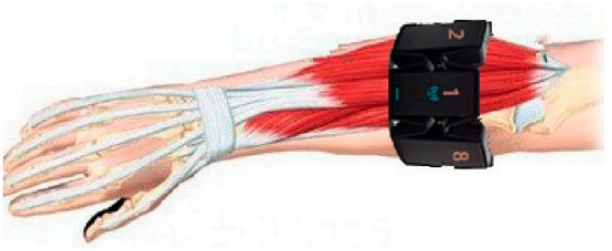

Figure 2. Wearing position and orientation.

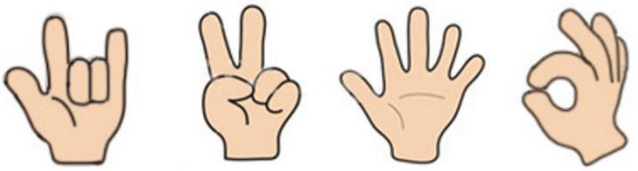

Figure 3. Four kinds of gestures collected in dataset-1.

4. Model and experiment description

4.1. Model

Unlike traditional machine learning algorithms, CNN models mainly include convolutional and pooling layers. In our proposed model, these two layers are both used.

Convolution is used to extract features, and by convolution layer, high-dimensional and effective features can be automatically extracted. The working principle is similar to convolution, but it is usually called cross-correlation in CNN. While realizing cross-correlation，the convolution window begins at the top leftmost corner of the input array and moves across it from left to right and top to bottom. There

DOI: 10.54254/2755-2721/4/2023305

are two important parameters in convolutional layer: padding and stride. Padding is the padding of elements (usually zeros) on either side of the input height and width. The number of rows and columns per slide in cross-correlation is called stride. These two parameters significantly affect the results of cross-correlation operation.

Pooling is mainly used to reduce the number of features extracted from the convolutional layer, to increase the robustness of the features or to reduce the dimensionality. Polling layer alleviates the over-sensitivity of the convolutional layer to position. Like convolutional layers, pooling layers obtain the output at a time by calculating elements in a fixed-shaped window (also known as a pooling window) of the input data. Unlike the convolutional layer where the mutual correlation between the input and the kernel is computed, the maximum or mean value of the elements in the pooling window is calculated directly by the pooling layer. These two operations are also referred to as maximum pooling or average pooling, respectively. The pooling window in max-pooling begins at the top leftmost corner of the input array and glides across the input array in left-to-right, top-to-bottom manner. The greatest value of the input subarray in the pooling window corresponds to the item at the specified position in the output array when the pooling window advances to a certain location. While in average pooling, the results become the average value of each pooling window.

4.2. Result analysis

For two datasets, we first applied machine learning classifier to train and validate data.

The average accuracy for dataset-1 using Random Forest can obtain 97%, which is quite high. Some other models, such as Decision Tree and AdaBoost were also used to compared with original option. The average accuracy for Decision Tree is 94%, and for AdaBoost is about 90%. In contrast, the accuracy for Random Forest has the highest accuracy. To get better results, the Convolutional Neural Network model was adopted to calculate the accuracy. It is surprising that its accuracy can reach 98% that is even higher than Random Forest.

For dataset- 2, we did the same work. However, the accuracy of Random Forest can only be 57%. In addition, the other machine learning models for it are not ideal either. The accuracy for Decision Tree is just 37%, and for AdaBoost is 24%. These models do not own exact performance to support the classification. As a result, we tried to use CNN to realize the same operation. At first, the model with eight Cov2D-layers resamples 60 points to ensure uniform input sample shape. The 10-fold-cross-validation is used for model to find the super parameter values that make the model generalization performance optimal. Dropout refers to randomly dropping neurons with a certain probability during the training process to solve overfitting issues. Fortunately, the average accuracy can reach nearly 72% with the original parameter set. It is clear that CNNs do outperform machine learning models, so we can use CNN model to cope with dataset-2. After that, we also desired to change the parameters to achieve better performance. Each change and the corresponding results are illustrated in Table 1.

Table 1. The content of the changes and the corresponding average accuracy.

Parameter	Value	Average accuracy
Original	Sample points=60, 10-fold, dropout (0.1), 8 Conv2D layers	71.5%
Sample points	80	70.7%
	100	70.5%
	120	70.3%
k-fold	4	69.4%
	12	69.6%
dropout	dropout (0.3)	70.9%
	dropout (0.5)	70.5%
Cov2D-layers	7	64.5%
	6	56.2%

DOI: 10.54254/2755-2721/4/2023305

With the probability of dropping neurons temporarily set as 0.1, the average accuracy can reach 71.5% in the first attempt. Afterwards, each parameter is changed to different value to calculate the average accuracy for each option. Compared with original model, we obtained that the average accuracy of the model does not change considerably while changing sample points, k-fold, dropout and Cov2D-layers. The results of these models are quite closed to original model. However, the accuracy drops sharply when decreasing the number of Cov2D-layers, which means that the structure of CNN models plays a significant role in final results.

5. Conclusion

A classifier model based on CNN with application to EMG-based and smartphone-based datasets was put forward in this research. Compared with traditional machine learning algorithms, our model achieved a higher classification accuracy. The results of these two datasets were about 97% and 72%, respectively. To study the influence of different factors on the recognition performance of CNN, we also changed some parameters and the convolutional layer's number of the initial model. It was found that the structure of CNN models had the greatest impact on results.

In the future, further improvements can be made, such as collecting larger datasets for the validation of CNN models or trying more complex classification models, e.g., graph neural networks which have been successfully used in a series of problems [20][21]. We will also incorporate several structures tested for larger datasets for further application in real-world systems.

References

[1] Van Kasteren, T.L.M.; Englebienne, G.; Kröse, B.J.A. An activity monitoring system for elderly care using generative and discriminative models. Pers. Ubiquitous Comput. 2010, 14, 489–498.

[2] Rialle, V.; Duchêne, F.; Noury, N.; Bajolle, L.; Demongeot, J. Health "Smart" Home: Information technology for patients at home. Telemed. e-Health 2002, 8, 395–409.

[3] Fiorini, L.; Bonaccorsi, M.; Betti, S.; Dario, P.; Cavallo, F. Ambient Assisted Living. Ital. Forum Ambient Assisted Living 2016, 426, 251.

[4] Arif, S. , Jing, W. , Hassan, T. U. , & Fei, Z. . (2019). 3d-cnn-based fused feature maps with lstm applied to action recognition. Future Internet, 11(2), 42.

[5] Wang, J.; Hu, F.; Li, L. Deep Bi-directional long short-term memory model for short-term traffic flow prediction. Lect. Notes Comput. Sci. 2017, 306–316.

[6] Yu, Z. , Zhao, J. , Wang, Y. , He, L. , & Wang, S. . (2021). Surface emg-based instantaneous hand gesture recognition using convolutional neural network with the transfer learning method. Sensors, 21(7), 2540.

[7] Parvathy, P., Subramaniam, K., Prasanna Venkatesan, G. K. D., Karthikaikumar, P., Varghese, J., & Jayasankar, T. (2021). Development of hand gesture recognition system using machine learning. Journal of Ambient Intelligence and Humanized Computing, 12(6), 6793-6800.

[8] F. Peng, C. Chen, X. Zhang, X. Wang, C. Wang and L. Wang, "sEMG-based Gesture Recognition by Rotation Forest-Based Extreme Learning Machine," 2021 IEEE International Conference on Real-time Computing and Robotics (RCAR), 2021, pp. 1122-1127, doi: 10.1109/RCAR52367.2021.9517479.

[9] Jiang, W. (2021). Applications of deep learning in stock market prediction: recent progress. Expert Systems with Applications, 184, 115537.

[10] Jiang, W., & Zhang, L. (2018). Geospatial data to images: A deep-learning framework for traffic forecasting. Tsinghua Science and Technology, 24(1), 52-64.

[11] Jiang, W., & Zhang, L. (2020). Edge-siamnet and edge-triplenet: New deep learning models for handwritten numeral recognition. IEICE Transactions on Information and Systems, 103(3), 720-723.

[12] Asadi-Aghbolaghi, M., Clapes, A., Bellantonio, M., Escalante, H. J., Ponce-López, V., Baró, X., ... & Escalera, S. (2017, May). A survey on deep learning based approaches for action and gesture

DOI: 10.54254/2755-2721/4/2023305

recognition in image sequences. In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017) (pp. 476-483). IEEE.

[13] Oyedotun, O. K., & Khashman, A. (2017). Deep learning in vision-based static hand gesture recognition. Neural Computing and Applications, 28(12), 3941-3951.

[14] Yang, K., Xu, M., Yang, X., Yang, R., & Chen, Y. (2021). A Novel EMG-Based Hand Gesture Recognition Framework Based on Multivariate Variational Mode Decomposition. Sensors, 21(21), 7002. https://doi.org/10.3390/s21217002

[15] Zdemir, M. A. , Kisa, D. H. , O Güren, Onan, A. , & Akan, A. . (2020). EMG based Hand Gesture Recognition using Deep Learning. 2020 Medical Technologies Congress (TIPTEKNO).

[16] Asif, A. R., Waris, A., Gilani, S. O., Jamil, M., Ashraf, H., Shafique, M., & Niazi, I. K. (2020). Performance Evaluation of Convolutional Neural Network for Hand Gesture Recognition Using EMG. Sensors, 20(6), 1642. https://doi.org/10.3390/s20061642

[17] Panella, M. , & Altilio, R. . (2018). A smartphone-based application using machine learning for gesture recognition: using feature extraction and template matching via hu image moments to recognize gestures. IEEE Consumer Electronics Magazine, 8(1), 25-29.

[18] Thapa, K., Abdullah Al, Z. Md., Lamichhane, B., & Yang, S.-H. (2020). A Deep Machine Learning Method for Concurrent and Interleaved Human Activity Recognition. Sensors, 20(20), 5770. https://doi.org/10.3390/s20205770

[19] Kim, J. H. , Hong, G. S. , Kim, B. G. , & Dogra, D. P. . (2018). Deepgesture: deep learning-based gesture recognition scheme using motion sensors. Displays, 55(DEC.), 38-45.

[20] Jiang, W. (2022). Graph-based deep learning for communication networks: A survey. Computer Communications.

[21] Jiang, W., & Luo, J. (2021). Graph neural network for traffic forecasting: A survey. arXiv preprint arXiv:2101.11174.