J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

CNN-based Gesture Recognition using Motion History Image☆

Youjin Koh[1] Taewon Kim[1] Min Hong[2] Yoo-Joo Choi[1]

ABSTRACT

In this paper, we present a CNN-based gesture recognition approach which reduces the memory burden of input data. Most of the neural network-based gesture recognition methods have used a sequence of frame images as input data, which cause a memory burden problem. We use a motion history image in order to define a meaningful gesture. The motion history image is a grayscale image into which the temporal motion information is collapsed by synthesizing silhouette images of a user during the period of one meaningful gesture. In this paper, we first summarize the previous traditional approaches and neural network-based approaches for gesture recognition. Then we explain the data preprocessing procedure for making the motion history image and the neural network architecture with three convolution layers for recognizing the meaningful gestures. In the experiments, we trained five types of gestures, namely those for charging power, shooting left,shooting right, kicking left, and kicking right. The accuracy of gesture recognition was measured by adjusting the number of filters in each layer in the proposed network. We use a grayscale image with 240 x 320 resolution which defines one meaningful gesture and achieved a gesture recognition accuracy of 98.24%.

☞keyword : Gesture recognition, neural network, convolutional neural network, motion history image

1. Introduction

Gestures are expressive, meaningful body motions involving physical movements of the fingers, hands, arms, head, face, or body[1]. Gesture recognition approaches have been studied for a long time in the field of computer vision and pattern recognition[1]. Recently, this research field has been evolved rapidly by neural network technology. Approaches based on deep neural network have outperformed “non-deep” state-of-the-art approaches[2] and been applied to various application domains[3,4].

In the gesture recognition approaches based on the neural network, one gesture is usually represented by 20~30 frame images which make mobile computing impossible due to the large memory burden. We can consider two points to develop a neural network model for gesture recognition which can be executed on the low-capacity system. First, we should design a lightweight neural network model that is not too wide or deep. However, in this case, the recognition accuracy can be deteriorated due to the shallow model structure if the meaningful data preprocessing is not provided. The second point is to reduce the size of the input data while minimizing the impact on recognition accuracy.

In this paper, we propose a novel method to reduce the size of the input data of a neural network model for gesture recognition. In the proposed method, a series of frame images representing one gesture are processed into one grayscale image called a motion history image and used as input data for learning meaningful gestures using the neural network. Moreover, we present the neural network model with just three convolutional layers and two fully connected layers. Then we conduct the experiments that adjust the number of filters and nodes in each layer of the proposed CNN model to determine the optimal model that shows the best recognition accuracy. As a result, we achieved a gesture recognition accuracy of 98.24% without high memory capacity through the proposed neural network model and the

1 Department of Media Technology and AI Software Engineering, Seoul Media Institute of Technology, Seoul, 07590, South Korea
2 Department of Computer Software Engineering, Soonchunhyang University, Asan, 31538, South Korea
* Corresponding author: Yoo-Joo Choi
[Received 6 June 2020, Reviewed 15 June 2020(R2 30 July 2020), Accepted 21 August 2020]
☆ This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF-2017R1D1A1B03035718) funded by the Ministry of Education.
☆ A preliminary version of this paper was presented at ICONI 2019.

data preprocessing for the motion history image.

2. Related works

In this section, we briefly summarize the previous gesture recognition approaches based on non-neural network and neural network. And the input data formats for previous neural network models are analyzed.

Typically, the meaning of a gesture can depend on the spatial information(where it occurs), pathic infor mation(the path it takes), symbolic information(the sign it makes), and affective information (its emotional quality)[1]. A lot of approaches have been studied to extract the spatial, pathic, symbolic, and affective information from the sensor or image data for gesture recognition.

2.1 Non-neural network approaches for gesture recognition

There are various non-neural network approaches for gesture recognition ranging from statistical recognition, such as HMM(Hidden Markov Models)[5,6] and PCA(Principal Component Analysis)[7], Kalman filterings[8], particle filtering[9,10] and condensation algorithms[11].

The HMM is a double stochastic process governed by an underlying Markov chain with a finite number of states (X) and a set of random functions(Y), each associated with one state[1]. The goal is to learn about X by observing Y.

PCA is mostly used as a tool in exploratory data analysis and for making predictive models. Principal components means basis vectors that are uncorrelated into different individual dimensions. PCA is either done by singular value decomposition of a design matrix or by calculating the data covariance matrix of the original data and performing eigenvalue decomposition on the covariance matrix[12].

In statiscs and control theory, Kalman filtering is an algorithm that uses a series of measurements observed over time, containing statistical noise and other inaccuracies, and produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone, by estimating a joint probability distribution over the variables for each timeframe[13].

Particle filters have been very effective in estimating the state of dynamic systems. The key idea is to represent probability densities by set of samples. It has the ability to represent a wide range of probability densities, allowing real-time estimation of nonlinear, non-Gaussian dynamic systems[1].

The condensation algorithms have been proposed[11, 14] to automatically switch between various prediction motion models by using multiple models to predict different types of motion of the objects. They significantly improved the performance of the tracker. In addition, studies are being conducted to overcome the limitations of the existing methods[15].

2.2 Neural network approaches for gesture recognition

An artificial neural network is composed of artificial neurons or nodes for solving artificial intelligence problems. In general, the neural network which is composed only of the fully connected layers, is likely to cause spatial information loss. Therefore, Convolutional Neural Network (CNN) has been proposed to solve the problem of loss of spatial information of data by maintaining the shape of input/output data of each layer. CNN effectively recognizes and extracts features from each frame image in the learning process. The CNN usually includes the convolution layer, pooling layer and fully-connected layer. The convolutional layer is a core component of the CNN, and the set of learnable filters varies according to the type and depth of each layer, and accordingly, the result of learning also varies[16]. Through the pooling layer, it is possible to prevent an excessive increase in the amount of processed data. The fully connected layer is the same as the traditional multi-layer perceptron neural network. The CNN has been widely applied to object classification and static gesture (pose) recognition.

The temporal dimension in sequences typically causes gesture recognition to be a challenging problem in terms of both amounts of data to be processed and model complexity[2]. 3D CNN approaches have been proposed for spatio-temporal feature extraction[17]. Molchanov et al. extended a 3D CNN with a recurrent mechanism for detection and classification of dynamic hand gestures[18]. In

[18], A 3D CNN for spatio-temporal feature extraction is followed by a recurrent layer for global temporal modeling and a softmax layer for predicting class-conditional gesture probabilities.

The long short-term memory(LSTM)[19] cells for RNN(Recurrent Neural Networks) was introduced by Hochreiter et al. in 1997. LSTMs are an important part of deep models for image sequence modeling for gesture recognition[2, 20].

The new approaches that utilize a double flow structure has emerged for more efficient gesture recognition. J. Duan et al. proposed a convolutional two stream consensus voting network(2SCVN) which explicity models both the short-term and long-term structure of the RGB sequences[21]. In their approaches, a 3D depth-saliency ConvNet stream (3DDSN) was aggregated in parallel to identify subtle motion characteristics. These two components in an unified framework significantly improve the recognition accuracy. Weinzaepfel et al. [22] proposed a novel approach in which dense optical flow maps or iDTs detects frame proposals and scores them with a combination of static and motion CNN features for action localization. Simonyan et al. [23] proposed another two-stream convolutional network. In their approach, to complement the dynamic information along to the temporal flow in addition to static information along to the spatial flow of the input data, optical flow processed data was used as learning input data. Then, the feature map output from each neural network was fused to improve recognition accuracy.

2.3 Input data format for deep learning based approaches

Input data preprocessing is also important to achieve good results in gesture recognition research. There are various image datasets for each gesture recognition study. Among them, ChaLearn LAP, VIVA Challenge Dataset, and UCF-101, UCF-Sports which are based on YouTube videos, are continuously used as the original dataset for research on gesture recognition. These data sets include a large amount of raw data and data preprocessing is required for each learning approach. Table 1 depicts the datasets and the input data format that previous gesture recognition methods [17,18,21,22] have used. These methods used 25-80 images for an input gesture data.

(Table 1) The input datasets of previous gesture recognition methods

Year(ref.)	2015(17)	2016(18)	2016(21)	2015(22)
network	3DCNN	R3DCNN	2SCNN 3DDSN	Spat+motion CNN
Used dataset	VIVA challenge	ChaLearn, VIVA challenge	ChaLearn IsoGD	UCF-101
The number of input image frames for one gesture	32	80	32	25
Image size	57 x 25	112 x112	32 x 32	227x227
modality	RGB	RGB, D	RGB, D	RGB
accuracy	77.50	98.20	96.74	54.28

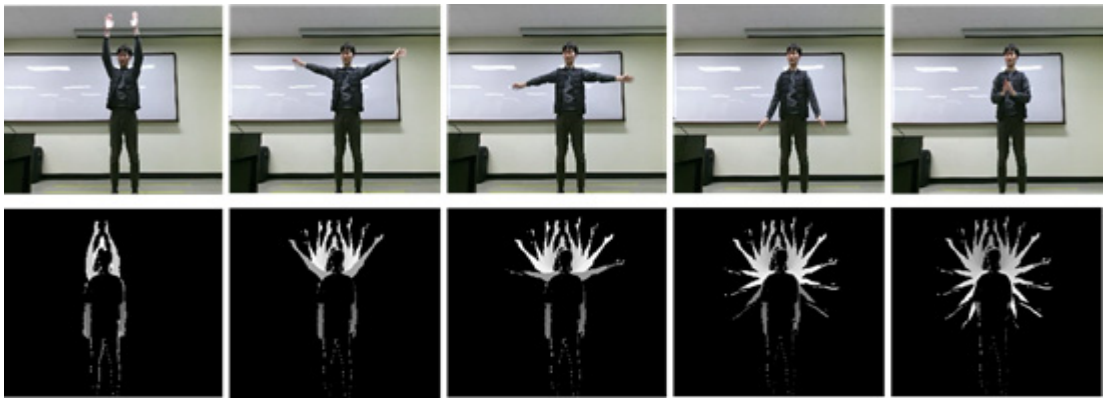
3. Proposal Method

3.1 Motion history image

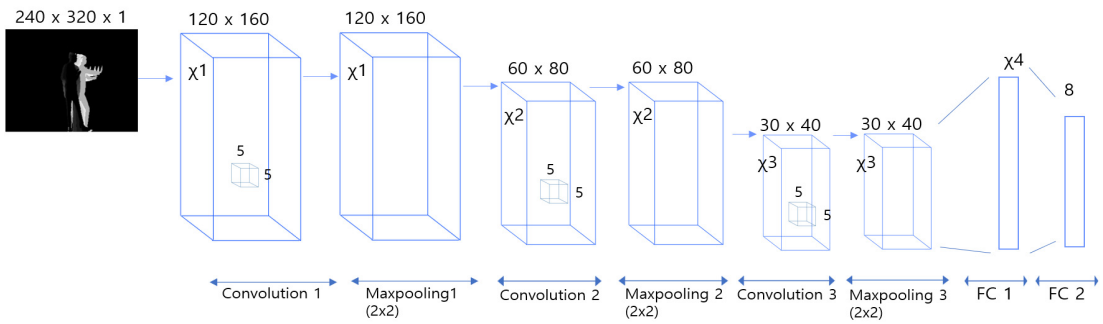
In general, a single gesture is represented by dozens of frame images. However, our method uses a motion history image to represent a gesture. A motion history image is a static image that includes the body moving path for the representation of a gesture. It can be used to effectively represent a single gesture in a small memory. A sequence of body silhouettes for a gesture is compressed into a gray scale image in which the main motion information is preserved. It can be used to more effectively represent the flow or sequence of motion than a color images. The proposed method classifies the types of gestures using single motion history image as input to the CNN. Figure 1 sequentially illustrates a series of motion history images for the power-charging gesture. All motion history images in the sequence are used for training the gesture.

3.2 Three-layered CNN

Figure 2 depicts the structure of the proposed CNN. The model has a simple structure including just three convolution layers with two max pooling layers, and two fully connected layers. In each convolution layer, 5x5 filter is applied. A



(Figure 1) A sequence of motion history images for a power-charging gesture



(Figure 2) Three-layered Convolutional Neural Network

grayscale image with 320 x 240 resolution is used as input data for the gesture. And we measured the recognition accuracy by adjusting the number of filters in each convolution layer and the number of nodes in the fully connected layers to determine the optimal CNN. Eight gestures including five meaningful gestures and three meaningless movements have been trained using the proposed network.

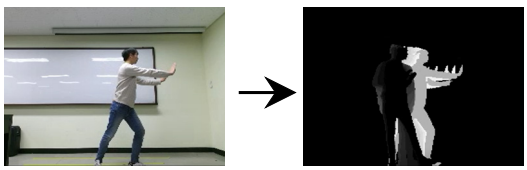
4. Experiment

4.1 Experiment environment

We asked seven participants to perform five different dynamic gestures. Figure 3 shows a color image and the selected motion history image for 'shooting left' gesture. Figure 4 shows the motion history images for five gestures: 'power charging', 'shooting left', 'shooting right', 'kicking left', 'kicking right'. Figure 5 depicts the motion history images of three error types. The error types mean the motion history images for abnormal user movement, not target gestures. The first error type is that the motion history image is completely black because there is no user's movement. The second error type is that most area of the motion history image is defined as a user area due to the user's movement at a distance close to the camera. The third error type is that the motion history image contains a small white area that looks like noise due to the user's fine movements before the target gesture.

We obtained 3644 motion history images for training and 1723 motion history images for test data from seven participants. Training data and test data are in a ratio 7:3.

Table 2 shows the dataset for each gesture used in our gesture learning experiments. The gesture learning and inferring were conducted in the experimental environment shown in Table 3.



(Figure 3) A color image and a motion history image of the 'shooting left' gesture

(a)shooting right (b)shooting left (c) kicking right

(d) kicking left (e) power charging

(Figure 4) Selected motion history images for five target gestures.

(a)shooting right (b)shooting left (c) kicking right

(Figure 5) The motion history images of three error types.

(Table 2) Number of motion history images used in gesture learning and inferring

Gesture type	Training	Test
Power- charging	611	261
Shooting left	613	333
Shooting right	613	333
Kicking left	555	253
Kicking right	554	246
Error-Black	241	102
Error-Etc	318	136
Error-Ready	139	59
Total	3644	1723

(Table 3) Number of filters in Convolution Layer and Fully connected layer nodes

OS	Windows10
CPU	intel Core i7-6700K @ 4.0GHz RAM 64.0GB
GPU	NVIDIA GeForce GTX 1080 Ti

4.2 Experimental results

We conducted a gray-scale MHI-based gesture recognition experiment using the proposed CNN and motion history image datasets. The experiment was repeated with the different number of filters of the convolution layers and with the different number of nodes of the fully connected layer.

(Table 4) Number of filters in convolution layers and number of nodes in fully connected layers

Name of layer	Conv1 (X1)	Conv2 (X2)	Conv3 (X3)	FC1 (X4)	FC2
Number of filters	32	32	32	32	8
	16	16	16	16	8
	8	8	8	8	8

*X1-X3: Number of filters in conv1, conv2 and conv3 layers
*X4 : Number of nodes in fully connected layer

Table 4 shows the number of filters and nodes of the proposed CNN. In the experiments, the learning rate was set to 0.001 and the number of epochs was set to 300. We also applied the learning rate decay to improve the accuracy. Table 5 shows the recognition accuracy according to application of decay. In the gesture learning, the application of learning rate decay showed higher accuracy, and the runtime was almost identical in both cases. Finally, we achieved a gesture recognition accuracy of 98.24% with eight filters in each convolution layer and 8 nodes in the fully connected layer.

(Table 5) Difference of recognition accuracy according to application of learning rate decay

Application of Decay	Number of filters (X1, X2, X3)	Accuracy
Applied from 250 epoch	32	94.73
	16	96.49
	8	98.24
none	32	92.98
	16	94.73
	8	96.49

5. Conclusions

In this paper, we proposed a method that defines one gesture as one gray-scale image to reduce the memory for learning and processing gestures. In addition, it was confirmed that data for meaningless motion can help improve the recognition accuracy of the target gestures. Through the proposed method, the user can define meaningful gestures as a small memory and can greatly reduce the memory capacity for learning and inferring various meaningful gestures based on deep learning. As a future study, we intend to construct a deeper network that enables mobile computing to increase recognition accuracy while reducing the memory and computation burden.

References

[1] S. Mitra and T. Acharya, “Gesture Recognition: A Survey.” IEEE, Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), Vol. 37, No. 3, pp. 311-324, 2007.

[2] M. Asadi-Aghbolaghi, A. Clapes and M. Bellantonio. “A survey on deep learning based approaches for action and gesture recognition in image sequences.” IEEE 12th International Conference on Automatic Face & Gesture Recognition, pp. 476-483, 2017. https://doi.org/10.1109/FG.2017.150

[3] C. L. Lisetti and D. J. Schiano. “Automatic classification of single facial images.” Pragmatics Cong., Vol. 8, pp. 185-235, 2000.

[4] P. N. Huu, Q. T. Minh and H. L. The, ″An ANN(Artificial Neural Network)-based gesture recognition algorithm for smart-home applications.″ KSII Transactions on Internet and Information Systems, Vol. 14, No. 5, pp. 1967-1983, 2020. https://doi.org/10.3837/tiis.2020.05.006

[5] L. R. Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition,” IEEE, Vol. 77, No. 2, pp. 257-285, 1989.

[6] J. Yamato, J. Ohya and K. Ishii, “Recognizing human action in time sequential images using hidden Markov model,” in Proc. IEEE, Conf. Comput. Vis. Pattern Recogn., Champaign, IL, pp. 379-385, 1992.

[7] S. Arulapalam, S. Maskell, N. Gordon and T. Clapp, “A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking,” IEEE, Trans. Signal Process., Vol. 50, No. 2, pp. 174-188, 2001.

[8] G. Welch and G. Bishop, “An introduction to the Kalman filter,” Dept. Comput. Sci., Univ. North Carolina, Chapel Hill, Tech. Rep. TR95041, 2000.

[9] S. Arulapalam, S. Maskell, N. Gordon and T. Clapp, “A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking,” IEEE Trans. Signal Process., Vol. 50, No. 2, pp. 174-188, 2001.

[10] C. Kwok, D. Fox and M. Meila, “Real-time particle filters,” Proc. IEEE, Vol. 92, No. 3, pp. 469-484, 2004.

[11] M. Isard and A. Blake, “Condensation-Conditional Density Propagation for Visual Tracking”, International Journal of Computer Vision, Vol. 29, No. 1, pp. 5-28, 1998.

[12] Principal component analysis, https://en.wikipedia.org/wiki/Principal_component_analysis

[13] Kalman filter, https://en.wikipedia.org/wiki/Kalman_filter

[14] M. Isard and A. Blake, “A mixed-state condensation tracker with automatic model-switching,” in Proc. 6 int. Conf. Comput. Vis., Mumbai, India, pp. 107-112, 1998.

[15] T. Pei, C. Guozhen and L. Nianfeng, ″Study on gesture recognition based on IIDTW(Improved Interpolation Dynamic Time Warping) algorithm.″ KSII Transactions on Internet and Information Systems, Vol. 13, No. 12, pp. 6063-6079, 2019. https://doi.org/10.3837/tiis.2019.12.015.

[16] Convolutional Neural Network, https://en.wikipedia.org/wiki/Convolutional_neural_network#Convolutional_layer

[17] P. Molchanov, S. Gupta, K. Kim and J. Kautz. “Hand gesture recognition with 3d convolutional neural networks.” CVPRW, 2015, pp. 1-7, 2015. https://doi.org/10.1109/CVPRW.2015.7301342

[18] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. “Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network.” CVPR, 2016. https://doi.org/ 10.1109/CVPR.2016.456

[19] S. Hochreiter and J. Schmidhuber. “Long short-term memory.”, Neural computation, Vol. 9, No. 8, pp.1735-1780, 1997.
[20] B. Singh, T.K. Marks, M. Jones, O. Tuzel, and M. Shao. “A multi-stream bi-directional recurrent neural network for fine-grained action detection.”, In CVPR, 2016.
[21] J. Duan, S. Zhou, J. Wan, X. Guo, and S. Z. Li. “Multi-modality fusion based on consensus-voting and 3d convolution for isolated gesture recognition.”, 2016. arXiv:1611.06689.
[22] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. “Learning to track for spatio-temporal action localization.”, abs/1506.01929, 2015. arXiv:1506.01929v2
[23] K. Simonyan and A. Zisserman. “Two-stream convolutional networks for action recognition in videos.” NIPS, pp. 568-576, 2014. https://doi.org/10.1109/MMSP.2018.8547088

◐ 저 자 소 개 ◑

고 유 진(You-Jin Koh)

2017 : KyungHee University(Korea), Department of Japanese language(Bachelor degree)
2018~2020 : Department of Media Technology(Master degree), Seoul Media Institute of Technology (SMIT)
Interests : Augmented Reality, Human-computer Interaction
E-mail : youjink@smit.kr

김 태 원(Tae-Won Kim)

2018 : Namseoul University(Korea), Department of Computer Science(Bachelor degree)
2017~2018 : ducogen.Co.,Ltd (Research Engineer)
2019~Present : Department of AI Engineering Software(Master degree), Seoul Media Institute of Technology (SMIT)
Interests: Computer Vision, Computer Graphics, Augmented Reality, Human-computer Interaction
E-mail : wingtgniw@smit.kr

홍 민(Min Hong)

1995: Soonchunhyang University (Korea), Computer Science (Bachelor degree)
2001: University of Colorado at Boulder (USA), Computer Science (Master degree)
2005: University of Colorado at HSC (USA), Bioinformatics (PhD degree)
2006~Present: Professor at Department of Computer Software Engineering, Soonchunhyang University, Korea
Interests: Computer Graphics, Dynamic Simulation, Bioinformatics, Image Processing
E-mail: mhong@sch.ac.kr

최 유 주(Yoo-Joo Choi)

1991: Ewha Womans University (Korea), Computer Science (Master degree)
2005: Ewha Womans University (Korea), Computer Science (PhD degree)
2006~Present: Professor at Department of AI Engineering Software, Seoul Media Institute of Technology (SMIT), Korea Research
Interests: Image Processing, Computer Vision, Computer Graphics, Augmented Reality, Human-computer Interaction
E-mail: yjchoi@smit.ac.kr