A Gesture Controlled Music Playback System Using Convolutional Neural Network

[1]Hung-Kuang Chen, *Member IEEE*, [2]Che-Chia Chuang
Electronic Engineering Department,
National Chin-Yi University of Technology,
Taichung, Taiwan, R.O.C.
[1]hankchentw@gmail.com, [2]tau852852@yahoo.com.tw

2020 International Symposium on Computer, Consumer and Control (IS3C) | 978-1-7281-9362-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/IS3C50286.2020.00010

***Abstract*—In this paper, we propose an application of applying hand gesture recognition using Convolutional Neural Network(CNN) for music playback control. Our system begins with live image capturing from a USB camera; subsequently, followed by the initialization, calibration, and motion detection stages. To train the neural network, a set of segmented and tagged hand gesture images were used. According to our experimental results, our method has the advantages of low latency and high accuracy, which is capable of detecting a set of six gestures in real-time.**

***Keywords*-hand gesture, Convolutional Neural Network, machine learning**

I. Introduction

Interactions between human and computers including speech, brain waves, gestures, and body movements have been studied for the last two decades. Human gestures were considered as an effective human-machine communication devices in addition to the keyboard, mouse, joystick, pads, and touch pads [1]. Being the most dexterous part of our body, hand gestures have long been used for communication in various situations. A great variety of techniques and devices such as mechanical trackers, bend resistors, optical fiber, accelerometers, etc., have been used in data glove devices for hand gestures tracking or recognition. However, most of these approaches were mainly restricted by the capability of tracking device and their costs.

As an alternative, vision based approaches from computer vision research for real-time gesture recognition were intensively studied [2]. Famous applications in computer gaming and virtual reality including Microsoft Kinect, Sony PlayStation Eye, Creative Senz3D, and Mesa Swiss-Ranger, etc., freed their users from remote controller by vision based gesture recognition. A number of surveys on the proposed techniques can be found at [3]–[5].

Previous works revealed several difficulties including complex background, partial occlusion, varying object scales, moving viewpoint, and varying lighting conditions, etc. To perform fast and effective feature extraction and high-precision low latency hand gesture recognition, traditional techniques usually failed. With the help of recent developments on Graphics and AI processors for Deep Learning applications, recent works suggested applying neural networks, especially the deep neural networks, to image or gesture recognition [6]–[24]. A more comprehensive survey has been done in [25].

In this paper, we proposed a vision-based hand gesture recognition approach based on convolutional neural networks(CNNs)and its application to intuitive music playback control. With our system, the users can issue playback commands by various hand gestures captured live from a USB camera. The captured images were calibrated, processed, and fed to CNNs for further classification. Commands of a music playback system were then issued by executing scripts corresponding to the classification results. In our application, six hand gestures were used by the music playback system. For each gesture, 1500 different images were manually tagged and used for training or testing: namely, 750 of them were used for training and the other 750 images were used for testing. According to our experimental results, the proposed method can successfully detect the six gestures at high precision rates and issue corresponding commands in real-time.

II. Convolutional Neural Networks

We adopted vision-based approach to hand gesture recognition in combination with the Convolutional Neural Network(CNN). Discussions on relate works and terminologies are therefore restricted to those vision-based approaches to hand gesture recognition and the CNNs.

The Convolutional Neural Networks, denoted as CNNs or ConvNets, is a kind of multilayer perceptron (MLP) neural network comprising three types of layers: namely, the convolution, pooling, and fully-connected layers. The internal layered structure of a set of CNNs must be defined according to its application and is usually determined by the convolution filters, size and number of pools, pooling players, number of layers, and choice of activation function. A number of well-known architectures proposed so far include LeNet [26], AlexNet [27], VGGNet [28], GoogLeNet [29], ResNet [30], DenseNet [31].

III. CNN based hand gesture recognition and its application to music playback control

The discussions of our method and its application start with the motion detection of live capturing of hand gestures.

978-1-7281-9362-5/20/$31.00 ©2020 IEEE
DOI 10.1109/IS3C50286.2020.00010

Subsequently, the training and testing of the CNNs. Finally, the integration with the music playback system is explained.

A. Motion Detection and Live Capturing of Hand Gestures

The input is initially a sequence of images captured live from a USB camera. We begin with an initialization of background images; then, substrate the background values from the subsequently captured images and give a grey thresholded image. To reduce the interference of noises, the thresholded images can be binarized as shown in Figure 1.

Figure 1. An example of threshold image(left) and its binarization(right).

B. Construction of Training and Testing Data-sets

As shown in Figure 2, we have adopted six hand gestures for music playback control. For each gesture, 750 distinct images were used for training and another 750 distinct images were used for testing. Totally, an amount of $1,500 \times 6 = 9,000$ distinct images were used.

C. Training of CNNs

The architecture of the CNNs adopted by our system is shown in Figure 3. The output of the first and second convolution layer are 32 and 64, respectively. Each convolution layer is connected to a max pooling layer to avoid dense computations. Subsequent to the flatten layer, all the data were linearized and fed to a hidden layer of 64 nodes activated by Relu. Since the total number of nodes, $128 \times 128 \times 64 = 262,144$ nodes, required by the hidden layer is too large, we adopted the dropout approach suggested by [32] and save 50% of nodes to prevent from over-fitting.

The final decisions are made by the six output nodes activated by the Softmax function giving probability values ranging from 0 to 1. The training process is carried out individually for the six gestures. For each gesture, the parameters trained CNNs were exported to their corresponding folders. Assuming that the input of the Softmax function are $\mathbf{y} = y_1, y_2, \cdots, y_n$ then

$$\sigma(\mathbf{y})_i = \frac{e^{y_i}}{\sum_{k=1}^{n} e^{y_k}}, \text{for } i = 1, \cdots, n. \tag{1}$$

The finale decision is then made by choosing the node with maximum value. For example, if node 5 has the maximum value of 0.9 then gesture number 5 is recognized and the corresponding script are executed.

D. The music playback system

After the CNNs were properly trained, they can be integrated to a music playback system. Figure 4 shows a simple user interface of such music playback system where playback control commands such as the album selection, playback, pause, previous, and next, etc. are implemented.

For convenience of our discussion, we only take three music files named 01, 02, and 03, respectively. To carried out basic playback controls such as album selection, pause, play, previous, and next, etc. The gestures one, two, three correspond to the selection of album one(left), two(middle), three(right), respectively. We further distinguished the events according to the motion path of individual gestures and triggered the following events in Table .

Table I
THE PLAYBACK EVENTS.

Status	Gesture	Event
music stopped	0	-
music stopped	1	play
music stopped	2	-
music stopped	3	-
music stopped	4	-
music stopped	5	-
music stopped	move left	album 0(left)
music stopped	move right	album 2(right)
music playing	0	-
music playing	1	play
music playing	2	-
music playing	3	-
music playing	4	-
music playing	5	pause or continue
music stopped	move left	previous song
music stopped	move right	next song

IV. Experimental Results

Our experiments were performed on a PC with nVIDIA GTX 980Ti running Microsoft Windows 10. All the programs are coded in Python 3.7 with Google Tensorflow and Keras modules installed. For each gesture, 750 tagged images are used for training and 750 images for testing. The experiments were conducted on threshold and its binarized version independently on 50 layers with 100 times iteration for each layer. The results are listed in Table II.

Table II
THE TESTING RESULTS.

Training	Testing	Loss	Accuracy
binary	binary	0.1884	0.9740
threshold	threshold	0.0334	0.9887

V. Concluding Remarks and Future Works

In this paper, we proposed a hand gesture recognition method based on preliminary image analysis for hand motion detection and convolutional neural networks for various

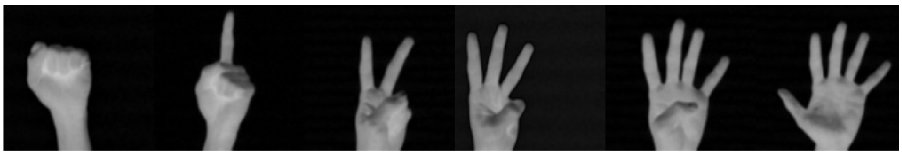

Figure 2. The six hand gestures for our music playback system.

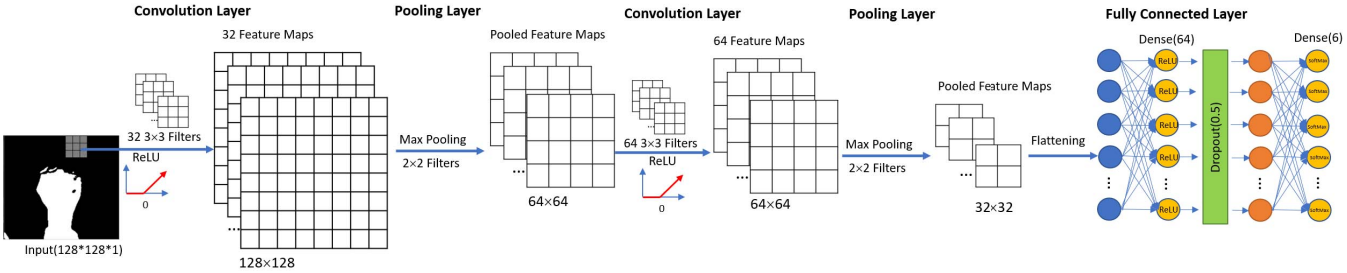

Figure 3. The architecture of the CNNs adopted by our system.

Figure 4. The user interface our music playback system.

hand gesture identification. As an example application, such method was applied to the control of a music playback system. According to our experimental results, the training time is low and the detection of various hand gestures can be done in realtime.

However, for a complex background, interferences were identified for noisy image with specular highlights or flashing lights; specifically, when the threshold of the front and back scenes is lower than 25 or when the threshold of the light source is unstable and the threshold is higher than 25. In the future works, we would like to work out a way to removing background interferences so that this system can be effectively executed in any environment.

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] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, pp. 1–43, Apr. 2011.

[3] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," *Computer Vision and Image Understanding*, vol. 141, pp. 152–165, 2015.

[4] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 1–54, 2015.

[5] S. Anwar, S. K. Sinha, S. Vivek, and V. Ashank, "Hand gesture recognition: a survey," in *Nanoelectronics, Circuits and Communication Systems*, pp. 365–371, Springer Singapore, 2019.

[6] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Static hand gesture recognition using artificial neural network," *Journal of Image and Graphics*, vol. 1, no. 1, pp. 34–38, 2013.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, p. I–647–I–655, JMLR.org, 2014.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1–7, 2015.

[11] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Static hand gesture recognition using principal component analysis combined with artificial neural network," *Journal of Automation and Control Engineering*, pp. 40–45, 01 2015.

[12] M. K. Ahuja and A. Singh, "Static vision based hand gesture recognition using principal component analysis," in *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*, pp. 402–406, 2015.

[13] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3593–3601, 2016.

[14] M. Han, J. Chen, L. Li, and Y. Chang, "Visual hand gesture recognition with convolution neural network," in *2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 287–291, IEEE, 2016.

[15] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Computing and Applications*, vol. 28, p. 3941–3951, Dec. 2017.

[16] F. Sultana, A. Sufian, and P. Dutta, "Advancements in image classification using convolutional neural network," in *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pp. 122–129, IEEE, 2018.

[17] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3d hand pose estimation from single depth images using multi-view cnns," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4422–4436, 2018.

[18] K. He, Z. Yang, L. Zhuang, and X. Zheng, "Hand gesture recognition based on convolutional neural network using a bistatic radar system," in *Eleventh International Conference on Signal Processing Systems*, vol. 11384, p. 113840Q, International Society for Optics and Photonics, 2019.

[19] G. Li, H. Tang, Y. Sun, J. Kong, G. Jiang, D. Jiang, B. Tao, S. Xu, and H. Liu, "Hand gesture recognition based on convolution neural network," *Cluster Computing*, vol. 22, no. 2, pp. 2719–2729, 2019.

[20] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, "Hand-gesture recognition using two-antenna doppler radar with deep convolutional neural networks," *IEEE Sensors Journal*, vol. 19, no. 8, pp. 3041–3048, 2019.

[21] Z. Zhang, K. Yang, J. Qian, and L. Zhang, "Real-time surface emg pattern recognition for hand gestures based on an artificial neural network," *Sensors*, vol. 19, no. 14, p. 3170, 2019.

[22] R. F. Pinto, C. D. Borges, A. Almeida, and I. C. Paula, "Static hand gesture recognition based on convolutional neural networks," *Journal of Electrical and Computer Engineering*, vol. 2019, 2019.

[23] F. Zhan, "Hand gesture recognition with convolution neural networks," in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 295–298, 2019.

[24] C.-M. Chang and D.-C. Tseng, "Loose hand gesture recognition using cnn," in *Advances in 3D Image and Graphics Representation, Analysis, Computing and Information Technology*, pp. 87–96, Springer, 2020.

[25] M. Asadi-Aghbolaghi, A. Clapés, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pp. 476–483, 2017.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[28] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv e-prints*, p. arXiv:1409.1556, Sept. 2014.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[31] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[32] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012.