Research Article
A Novel Deep Learning Technique for Higher Education System

Yali Tang\(^1\) and Lina Liu\(^2\)

\(^1\)Hebei Vocational University of Industry and Technology, Shijia Zhuang 050091, Hebei, China
\(^2\)Shijiazhuang Institute of Railway Technology, Shijia Zhuang 050062, Hebei, China

Correspondence should be addressed to Yali Tang; 20170111300075@stu.hubu.edu.cn

Received 19 January 2022; Revised 7 March 2022; Accepted 2 April 2022; Published 18 May 2022

Academic Editor: Naeem Jan

Copyright © 2022 Yali Tang and Lina Liu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This study aims to enhance the interest of university classrooms and the learning efficiency of students. From deep learning, gesture recognition algorithms (GRA) and long short-term memory (LSTM) are adopted to build prediction modules. After the system process is completed, a brand-new human-computer interaction (HCI) system is established. The deep learning GRA augments the operation of the recognition algorithm of the pattern matching method created on the previous image segmentation. Its enticements on two merits of gesture estimation are as follows: rapid detection of gesture joint features and convolutional neural network (CNN) image classification, thereby improving GRA. The recognition rates of “take”, “pinch”, and “point” are 98.6%, 99.5%, and 99.4%, respectively. When the data volume of the LSTM network model is less than 10, the prediction accuracy can reach 70%, and the performance is relatively stable. The designed HCI system can better recognize the teacher’s gesture intent in teaching, execute gesture commands, and optimize the interactive teaching method between the teacher and the computer. This system is a derivative form of gamification teaching. While it improves the interest in teaching, it is also of great significance for improving the teaching efficiency of general higher education.

1. Introduction

The continuous innovation of information technology has caused a lot of science and technology to change the way of teaching. In general, in higher education, the amount of knowledge imparted by teachers is often tremendous. The correlation is also robust between some subject knowledge. In the traditional teaching system, interaction has brought new learning methods for teachers and students. In the actual teaching process, the teacher will waste a lot of time because of the interaction of the teaching equipment. The development of artificial intelligence (AI) based on interactive teaching makes natural gestures to control multimedia equipment and provides ideas for reducing the difficulty of the teacher’s operation of the equipment. The gamification teaching methods have high requirements for gesture recognition (GR). The research optimization of GR is a critical topic in human-computer interaction (HCI).

GR has data-based gloves and vision-based recognition methods [1, 2]. Tseng et al. [3] split the gesture image from the background and then binarized it. The mathematical morphology of corrosion expansion and noise reduction processes the binarized gesture graphics. Before the GR, the previous binary image of the gesture was repeatedly corroded. Parts of the finger and palm images are segmented to show a different state of the finger images. The recognition principle is to analyze the number of scattered points first and then perform GR. Cheng et al. [4] used the skin color filter method to segment the hand image and then converted the image into black and white. The algorithm of the fingertip angle performs the calculation of gestures. The gesture image is cut and processed. An inappropriate background environment will affect image segmentation. In recognition, the angle of the gesture has a significant influence. With the same motion, the angle difference will harm the accuracy of praise, and the probability of misidentification will be high. In the field of HCI, Li [5], through the basic principle of a finger guessing game, used gesture estimation to effectively detect the gesture features of joints and applied CNN (a network that is used in image recognition and processing) to...
plan and process the joint features, which solved the bottleneck problem of the poor effect of gesture image segmentation in different environments. Among gesture-based HCI technologies, the essential thing is GR. At this stage, GR technology first processes images with red, green, and blue (RGB) and then uses the model to match [6]. This type of GR method has many shortcomings. The difficulty coefficient of gesture image segmentation is large in image preprocessing and complex backgrounds. In the GR based on model matching, the direction of the gesture will cause misrecognition, thus affecting the recognition rate. The speed will decrease segment recognition.

A new gesture recognition algorithm (GRA) is come forward by deep learning to address the shortcomings of GR. The algorithm draws on the relevant technology of GR at this stage and finds an algorithm different from the previous gesture estimation. The algorithm combines the pose estimation to find the standard features of the gesture quickly. CNN is used to classify images, and the speed and accuracy of recognition are improved. A deep learning HCI system is constructed based on gestures. The long short-term memory (LSTM) network predicts human gesture behavior. Through experiments, the predictive ability has been demonstrated of the interactive system. The designed HCI system can better recognize the meaning of gestures in the teacher’s teaching process and carry out the teacher’s gesture commands to optimize the interactive teaching method between the teacher and the computer.

The arrangement of the paper is as follows: Section 2 explains the methods used in the paper to get results. Section 3 discusses the results obtained by applying methods. Section 4 elaborates the conclusion of the paper.

2. Materials and Methods

In this section, the algorithm of gamification is well-explained, and a brief study on the storage of gesture images is also discussed. Then, this elaborates on the module of GR, the module of gesture prediction, and the design of the HCI system. Finally, the experiment to get verified results is done.

2.1. Application of Gamification Education. The gamification element is an extraordinary element in the game, which can be extracted and used in the game. It is a key component of gameplay. Gamification elements can also be specific game tools. Different gamification elements are different tools in the toolbox. Different game elements produce different effects. There are many gamification elements in the game, and the definitions of various gamification elements are also quite different. The types of gamification elements are shown in Figure 1.

In Figure 1, in interacting with computing, teachers use more gestures to control the computer. Therefore, visual elements are the critical elements used in gamification education. HCI in the teacher’s classroom is shown in Figure 2.

Machine learning (ML) technology is continuously optimized, and data are used to predict human behavior. For a human, the more experienced he is, the greater the chance of correct judgment he will have. There is a positive correlation between the two. ML adds new impetus to the design of new HCI systems. In gesture interaction, Mora games are considered to be the easiest game of interaction. In Mora games, players can use GR to express their intentions and identify each other’s actions. During the teacher’s teaching process, the computer executes the teacher’s behavioral instructions.

The simulated AI robot NAO (Biped humanoid intelligent robot) is used to build an HCI system for the Mora game. The human and the machine are interacted with a Mora game. The system consists of two parts: GR and the prediction module. It achieves the purpose of predicting the opponent’s motions based on the past data of the player’s gestures each time they play the game. NAO robots, GR, and prediction algorithms with insufficient computing power must be operated on a back-end processing platform. GR and the determination of victory or defeat are performed at the workstation. Then, the prediction module is called to estimate the probability of the person’s next punching gesture. An action command is issued to the NAO robot to perform the punching action. The system architecture diagram is shown in Figure 3.

2.2. Storage of Gesture Images. If the image captured by the camera is saved in an electronic device, the optical image must be converted into a digital image. Therefore, mastering the storage method of digital images is an essential condition of image processing. Under normal circumstances, there are two types of digital image storage: bitmap storage and vector storage [7]. Vector graphics are generally stored in geometric shapes and text fields, not corresponding to image objects. Therefore, here, the bitmap storage method will be discussed in detail. Bitmap storage generally discretizes the image. The image will be divided into many pixels. The color data are processed separately by channels. The image will be retained in the disk as a matrix. The number of rows and columns of the matrix represents the pixel length and width of the enclosed image. The corresponding coordinate data represent the pixel value of the image. There are differences in the quality of pictures from different sources, and the number of pixel values is also different. The three-color resolutions of 8 bit, 16 bit, and 24 bit are used more often [8].

In real life, eight people are more likely to contact them. It can discretize the color into a number between [0, 25]. The numbers in this range are integers. In general, pixel images are divided into two types based on channels. One is a single-channel grayscale image, and the other is a multichannel color image [9]. In the storage method of a single-channel grayscale image, each row and column coordinate of the image matrix will correspond to the corresponding pixel value. The size of the pixel value represents the depth of the image itself. For example, when the color resolution is 8 bits, the discrete color interval is [0, 25]. The smaller the color dispersion, the more pronounced the blackness of the image itself. The greater the color dispersion, the whiter the image itself. According to the numerical ratio, the intermediate value is adopted as the gradual gray value. In a multichannel
color image, if the image color resolution is 8 bits, three channels, the coordinate values of each row and column of the image matrix correspond to each other. The three-pixel values correspond to the red, green, and blue-gray values. For example, channel 0 means no blue, and 255 represents the bluest. Therefore, the image’s overall color is the combined effect of the gray values of the three channels [10].

Image storage is the basis of digital image processing. Fourier transform transforms the signal from the time domain to the frequency domain during signal analysis and processing. This is to analyze the signal frequency composition more accurately and lay the foundation for filtering operations. In image processing, after the Fourier transform is performed on an image, the image is transferred from the spatial domain to the frequency domain [11]. Fourier transform is the basis for image feature extraction. Image edge detection and filtering, and noise reduction are necessary conditions. If the pixel size of an image is specified as $M \times N$, the two-dimensional Fourier transform is shown as follows:

$$F(\mu, \theta) = \sum_{x=0}^{M-1} \sum_{y=1}^{N-1} f(x, y)e^{-i2\pi((\mu x/M)+(\theta y/N))}. \quad (1)$$

The $f(x, y)$ represents the rows and columns of pixels in the spatial domain. $F(\mu, \theta)$ represents the processed frequency domain value, expressed in a complex number. The converted frequency spectrum contains all the information of the image. The amplitude spectrum is used [12]. If $R$ and $I$ are used to represent the real part and need to be complemented in $F(\mu, \theta)$, then the amplitude spectrum of the image is shown as follows:
After the amplitude spectrum is acquired, the frequency of the image is displayed on the complex coordinates. In digital images, the spatial frequency represents the frequency of periodic changes in the gray image level within the relevant distance. It reflects whether the pixel value changes significantly, that is, the pixel value gradient in the plane space. When analyzing the amplitude spectrum, if there are more high frequencies, this indicates that the overall fineness of the image is positively correlated with the interference component [13].

\[
|F(\mu, \theta)| = \sqrt{R(\mu, \theta)^2 + I(\mu, \theta)^2}. \tag{2}
\]

2.3. Module for GR. The enhanced CPM (ECPM) algorithm for GR is proposed based on convolutional pose machines (CPM) [14, 15]. The CPM subnetwork and the identification subnetwork jointly construct the ECPM. The CPM subnetwork first quickly detects key features of gestures. Then, there is a skeleton block diagram of gesture features. After that, the feature map is transmitted to the recognition network. The skeleton images of the detected gesture features are classified. ECPM utilizes an end-to-end model, eliminating the need for gesture image segmentation and skin color detection. The network structure is shown in Figure 4.

There are many stages in the CPM network. Its first stage is five 9 \times 9 convolutional layers and two 1 \times 1 convolutional layers. The specific position of the joint is detected in the gesture image. The confidence map of the P + 1 layer is drawn through the fully connected layer. Thus, the confidence map is used to predict the output of other joints. One layer only represents the output of one joint and a background output [16, 17]. The composition is similar to the second stage to the fifth stage. The original image is connected to the confidence map of the previous location through a network composed of convolutional layers. Then, the structure is composed of 3 11 \times 11 convolutional layers and 2 1 \times 1 convolutional layers to form a relevant network. Thus, the P + 1 layer confidence map is output, and other joints’ result is inferred—all stages of CPM output heat maps of the orientation of all joints. The location is processed precisely. The final combined feature heat map is obtained [18, 19].

A loss function (a function that maps values of variables onto a real number) is applied to the output in the stage. The error is minimized between the predicted joint position and the ideal joint position. The optimal part of each common position is recorded as equations (3) and (4):

\[
b^p + (Yp) = Z, \tag{3}
\]

\[
f_t = \sum_{p=1}^{P+1} \sum_{z \in Z} \left| b^p_t(z) - b^p_t(z) \right|_2. \tag{3}
\]

Among them, \( p \) is the joint that goes through each place. \( z \) is the position of the corresponding joint. The sum of the loss function of each stage is shown in the following equation:

\[
F = \sum_{t=1}^{T} f_t. \tag{4}
\]

The network parameters of each stage are processed by the standard stochastic gradient descent method. The second and subsequent stages share the corresponding convolutional layer network weight values to obtain the result of sharing the image feature map [20, 21]. CNN is the last stage of ECPM. It is composed of 4 convolutional layers and 4 top pooling layers. Then, the features of the joint feature images output in the fifth stage are recorded through 3 fully connected layers. Finally, the output dimension of the fully connected layer is 3. It corresponds to 3 different gestures. Each gesture image is mapped to the type associated with it.

2.4. Module of Gesture Prediction. There is no pattern for players in a random mora in statistics. Therefore, the result of the mora cannot be accurately budgeted. When the data sample size of mora is large enough, it is processed and analyzed. The individual player’s mora sequence has specific characteristics of sequence. The pattern of this series of lines may become the theoretical statistical basis for predicting its mora [22, 23].

In a recurrent neural network (RNN), the improved LSTM is a prevalent model in deep learning. It is used to solve various problems. Under the premise of RNN, LSTM has more input gates, output gates, and forget gates. What’s more, the input and output are the cell state. It often affects the retained and forgotten information to a certain extent and makes up for the shortcomings of RNN’s long-term dependence. LSTM has shown the best state in completing data tasks in various sequences, including speech recognition and handwritten recognition [24, 25]. In recent years, LSTM has developed rapidly. It is widely used in analyzing time-series-related data. The LSTM network structure and the internal structure of a single cell are shown in Figure 5.

The sequence of human player actions is recorded as a time sequence \( A \{A_1, A_2, \ldots, A_n, \ldots\} \). In the data, the \( t \)-th step
is represented by $A_t$ in the action execution. Rock, scissors, and paper are represented by 0, 1, and 2, respectively. The LSTM network covers four hidden layers and one output layer. One is the value set by timestep. The output predicted value 0, 1, or 2 is the expected type of the next mora. The activation function is represented by tanh. The action sequence is used as a data set. The loss function uses mean-square error (MSE), as shown as follows:

$$\text{MSE}(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2,$$

where $\tilde{\theta}$ is the predicted value and $\theta$ is the actual value.

2.5. The Design of HCI System. There are three different gestures in the guessing game: scissors, paper, and rock. Paper wins rock, rock wins scissors, scissors win paper, and the tie is tied, corresponding to 4 kinds of results [26]. There are nine groups of all gesture result states of the guessing robot and the player. The interactive result is relatively simple. The robot is the decision node, the action is the state node, and the final result is the result node. The types of winning and losing combinations in Mora games are shown in Table 1.

The designed human-machine game system will be based on the NAO robot produced by Aldebaran Robotics. The collected image data are sent to the workstation. After the data are processed in the workstation, the control point is sent to the NAO humanoid robot to make the correct response action [27]. The NAO humanoid robot is 57 cm tall and has 25 degrees of freedom on the body. It can efficiently complete many kinds of actions. Two high-resolution cameras have wireless network image transmission, voice output functions, multiple languages, different platforms, and multilanguage programming. The highest resolution is 1280 * 960. The workstation is equipped with E5-2620 v4 dual CPU, 64 G memory, 500 G SSD, Quadro P5000 display adapter, and Ubuntu 18.04 operating system.

After the NAO is turned on, it will “reset” the system, connect to the host computer, and turn on the vision module. After this is done, the voice call invites other players to start guessing. Then, the time is calculated. During this period, it will run the trained LSTM human behavior to predict the model’s tendency toward human players and make a list of inferences. After the set time is up, the robot participates in the game with the result estimated by the algorithm. Human players do it at the same time. The NAO robot will call the ECPM algorithm to recognize the gesture when it obtains the image of the human gesture. Table 1 is used to determine the outcome of the outcome. All the player’s punch gestures will be stored in the “memory” and become the source of prediction data. The flow of the HCI system is shown in Figure 6.

In Figure 6, the design of the HCI system is explained completely. First, the voice invites the human player to play, invokes the behavior prediction model, and reacts accordingly to your prediction.

2.6. Experimental Setup. With the support of big data technology, a total of 2,000 gesture pictures are crawled on the website for experiments. A variety of static gestures are selected and organized. Among them, the static gesture of “take” is shown in Figure 7.

Gestures are roughly divided into three types: “pinch”, “take”, and “point”. After processing, 20% of the data set is used as the test data set, and 80% is used as the training data set for GR. To test the predictive performance of LSTM’s deep learning algorithm, two players played a game of guessing. Among them, the order of one person’s punches is recorded, and the total number of sets is 150. Rock, scissors, and paper are represented by 0, 1, and 2, respectively, forming a time series $\{0, 1, 2, \ldots\}$. The sequence of 100 times is regarded as the training set, the sequence of the last 50 times is regarded as the test set, and the training epoch is adjusted to 100.
Table 1: Combination types of wins and losses in Mora games.

|       | Scissors  | Stone    | Paper     |
|-------|-----------|----------|-----------|
| Scissors | A draw    | Robot wins | People win |
| Stone   | People win | A draw    | Robot wins |
| Paper   | Robot wins | People win | A draw    |

Figure 6: Flow chart of the system.

Figure 7: The gestures of static.

Figure 8: Comparison of GR rate.
3. Results

In this section, the GR is tested to confirm its accuracy rate, and the results are shown in Figure 8. Then, the prediction accuracy for authentic outcomes is analyzed.

3.1. Test of GR Accuracy. The GoogLeNet Inception V3 and Enhanced CPM algorithms are tested on the test data set. The result of the test is shown in Figure 8.

In Figure 8, the recognition rates of the Enhanced CPM algorithm in the three gestures of “take”, “pinch”, and “point” are 98.6%, 99.5%, and 99.4%, respectively. The recognition rate of GoogLeNet Inception V3 is 73.1%, 75.2%, and 80.2%. The recognition rate of actions that teachers need to use when presenting slides has been significantly improved.

3.2. Analysis of Prediction Accuracy. The drawing of the guessing result is shown in Figure 9.

In Figure 9, when the sample data volume is less than 10, the rate of design algorithm prediction accuracy is as high as 70%. However, as the number of data increases, the accuracy rate will tend to decline. After a slight fluctuation, the accuracy rate gradually rebounded and managed to a stable state. When the data volume is greater than 50, the accuracy rate is regular at 60%. The prediction accuracy rate on the training set reached 62.37%, and the prediction accuracy rate on the test set reached 64.44%. The results show that the designed predictive model has better performance.

4. Conclusion

With the development of AI, based on interactive teaching, the study of natural gesture-controlled multimedia devices provides ideas for reducing the difficulty of teachers’ equipment operation. The prediction module is constituted through GRA, robot NAO platform, and LSTM network from deep learning. After the system process is completed, a brand-new HCI system is created. Deep learning GRA is explored, and recognition algorithms based on image segmentation are optimized. This algorithm extracts the two advantages of gesture estimation, rapid detection of joint gesture features and CNN image classification, thereby improving GR accuracy. The guessing game completes an HCI system based on visual GR. The experiment confirms the algorithm’s performance used in the HCI system. However, the amount of sample data is low during the investigation of the experimental prediction module. The large probability will lead to the problem of small samples. In data prediction, the larger the data volume of the sample, the more convincing the experiment of the prediction module will be. The basic principle of relying on the guessing game is relatively simple. In the forthcoming years, the basic values will be further enriched in the development process, thereby improving the speed and quality of ML.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] X. Wang, Y. Wang, and Y. Qi, “Data exchange of relay transmission and information scheduling in shared networks,” IEEE Access, vol. 7, p. 1, 2019.
[2] X. Wang and K. Yan, “Immersive human-computer interactive virtual environment using large-scale display system,” Future Generation Computer Systems, vol. 96, pp. 649–659, 2017.
[3] K. T. Tseng, D. S. Chan, C. B. Yang, and S. F. Lo, “Efficient Merged Longest Common Subsequence Algorithms for Similar Sequences,” *Theoretical Computer Science*, vol. 708, pp. 75–90, 2018.

[4] M.-T. Cheng, L. Rosenheck, C.-Y. Lin, and E. Klopfer, “Analyzing gameplay data to inform feedback loops in the Radix Endeavor,” *Computers & Education*, vol. 111, pp. 60–73, 2017.

[5] J. Li, “Design of human-computer interaction system using gesture recognition algorithm from the perspective of machine learning,” in *Proceedings of the 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, IEEE, pp. 120–124, Fuzhou, China, September 2021.

[6] J. Wen, J. Zhang, and F. Wang, “Dynamic hand gesture recognition based on short-term sampling neural networks,” *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 01, pp. 114–124, 2021.

[7] J. Xie, X. He, L. Fang, and A. Plaza, “Scale-free convolutional neural network for remote sensing scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6916–6928, 2019.

[8] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, “Blind image quality assessment using a deep bilinear convolutional neural network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[9] Y. Li, G. Wang, L. Nie, Q. Wang, and W. Tan, “Distance metric optimization driven convolutional neural network for age invariant face recognition,” *Pattern Recognition*, vol. 75, pp. 51–62, 2018.

[10] S. Gai and Z. Bao, “New image denoising algorithm via improved deep convolutional neural network with perceptive loss,” *Expert Systems with Applications*, vol. 138, Article ID 112815, 2019.

[11] K. Yasaka, H. Akai, A. Kunimatsu, S. Kiryu, and O. Abe, “Deep learning with convolutional neural network in radiology,” *Japanese Journal of Radiology*, vol. 36, no. 4, pp. 257–272, 2018.

[12] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, “Multisource remote sensing data classification based on convolutional neural network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 937–949, 2017.

[13] W. Fang, L. Ding, B. Zhong, P. E. D. Love, and H. Luo, “Automated detection of workers and heavy equipment on construction sites: a convolutional neural network approach,” *Advanced Engineering Informatics*, vol. 37, pp. 139–149, 2018.

[14] S. Tammina, “Transfer learning using vgg-16 with deep convolutional neural network for classifying images,” *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, pp. 143–150, 2019.

[15] Q. Yuan, Q. Zhang, J. Li, H. Shen, and L. Zhang, “Hyperspectral image denoising employing a spatial–spectral deep residual convolutional neural network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1205–1218, 2018.

[16] S.-W. Chen and C.-S. Tao, “PolSAR image classification using polarimetric-feature-driven deep convolutional neural network,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 627–631, 2018.

[17] D. M. Pelt and J. A. Sethian, “A mixed-scale dense convolutional neural network for image analysis,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 2, pp. 254–259, 2018.