Hindawi
Complexity
Volume 2021, Article ID 6679746, 11 pages
https://doi.org/10.1155/2021/6679746

Research Article

Human Motion Gesture Recognition Based on Computer Vision

Rui Ma,[1,2] Zhendong Zhang ,[1] and Enqing Chen [3]

[1] *College of Physical Education (school Headquarters), ZhengZhou University, Zhengzhou 450001, China*
[2] *College of Physical Education, Pukyong University in South Korea, Busan 48513, Republic of Korea*
[3] *College of Information Engineering, ZhengZhou University, Zhengzhou 450001, China*

Correspondence should be addressed to Enqing Chen; ieeqchen@zzu.edu.cn

Received 10 November 2020; Revised 25 January 2021; Accepted 30 January 2021; Published 10 February 2021

Academic Editor: Wei Wang

Human motion gesture recognition is the most challenging research direction in the field of computer vision, and it is widely used in human-computer interaction, intelligent monitoring, virtual reality, human behaviour analysis, and other fields. This paper proposes a new type of deep convolutional generation confrontation network to recognize human motion pose. This method uses a deep convolutional stacked hourglass network to accurately extract the location of key joint points on the image. The generation and identification part of the network is designed to encode the first hierarchy (parent) and the second hierarchy (child) and show the spatial relationship of human body parts. The generator and the discriminator are designed as two parts in the network, and they are connected together in order to encode the possible relationship of appearance and, at the same time, the possibility of the existence of human body parts and the relationship between each part of the body and its parental part coding. In the image, the key nodes of the human body model and the general body posture can be identified more accurately. The method has been tested on different data sets. In most cases, the results obtained by the proposed method are better than those of other comparison methods.

1. Introduction

Human body gesture recognition is an important research direction of computer vision [1, 2]. Its ultimate purpose is to output the structural parameters of the person's overall or partial limbs, such as the outline of the human body, the position and orientation of the head, and the position of the human joint points or the category of parts. It should be said that the research methods of gesture recognition cover almost all theories and technologies in the field of computer vision, such as pattern recognition, machine learning, artificial intelligence, image graphics, and statistics [3, 4].

So far, many identification methods have been proposed. Dong et al. [5] proposed to learn the corresponding target contour model from the segmented image and then used the boost classifier to find the contour of the target in the image so as to obtain the position information of each part of the human body. The literature [6, 7] uses the HOG method to extract the information of each part of the human body in the image and then uses the classical algorithm support vector machine and random forest to identify and classify. Cui et al. [8] found the global optimal features from many features such as Fourier descriptors, shape context, edges, and gradients to quickly and accurately complete the back-projection process from features to three-dimensional poses. Zhang and Lu [9] used the histogram of gradient directions to restore the human pose and trained multiple local linear regressions to restore the human pose in a single frame of image. Li et al. [10] used the canny operator to extract edge features from the image in combination with pixel depth information and determined the head position of the person in the image through distance transformation and model matching and positioned the human body according to the prior human body proportion. Zhang et al. [11] used the iterated closest point (ICP) method to track the initialized human skeleton to achieve the purpose of human pose estimation. The literature [12] uses Markov random field (MRF) to segment the point cloud data containing the human body into four parts: head, torso, limbs, and background and recognizes the body's posture by part detection.

Nadeem et al. [13] first constructed a 3D grid model of the human body to find points of interest related to the measurement extreme value from the grid and then used these points to detect the head, hands, and feet of the human body.

In the past few years, research on human behaviour recognition based on depth images has received widespread attention [14–17]. Compared with ordinary optical images, because depth image pixels' record distance information and are color-independent, this overcomes the abovementioned problems encountered by ordinary optical images to a certain extent. With the development of optoelectronic technology, many researchers have combined the nature of depth images and applied many classic algorithms to such images. Alshawabkeh [18] combined pixel depth information and used canny operator to extract edge features from the image, through the distance transformation and model matching method, to determine the head position of the person in the image and locate the human body based on the prior human body proportion. Luo et al. [19] used computer graphics technology to construct a depth image database of human pose and used a classifier model to detect human body parts on a common PC. Jaffar et al. [20] used contour feature parameters combined with moments, directional gradient histogram, and human bone feature angles to perform multifeature fusion characterization of the depth information, color information, and bone information of the human body. Alzahrani et al. [21] comprehensively utilized the depth data and skeleton data provided by Kinect and effectively improved the real-time and robustness of gesture recognition through anthropometric knowledge and back propagation neural network. Ghazal et al. [22] used the Kinect camera to collect human bone information and 3D data corresponding to joint points for action recognition. Franco et al. [23] also used Kinect equipment to obtain depth information and designed a human posture recognition system specifically for sitting and standing postures.

This paper proposes a generative adversarial network to solve the problem of large deformations of parts of the body, while also considering the complexity of different levels of body parts. The internal structure of the generator and discriminator proposed in the article has been optimized, and it can simulate the hierarchical relationship between body parts. Hierarchical perception terms are also introduced in the objective function to standardize the relationship between parents and children. Hierarchical adversarial networks help to accurately estimate the positions of various parts of the body, especially body parts that are deformed or highly occluded.

2. Human Motion Gesture Recognition

According to the Oxford Dictionary, the posture of the human body is a special posture of the body and the way a person maintains his physical state [24]. Human posture recognition is the extraction, classification, and identification of human posture features and natural language description. It is a research hotspot that has received wide attention in recent years. It is a research on the application of human physiology, digital image processing, pattern recognition, and other disciplines field [25].

2.1. Technical Classification. From the perspective of the collection of human posture information, human posture recognition technology can be divided into two categories: contact recognition and noncontact recognition technology [26–29]. Contact recognition technology refers to a person's body wearing a sensor, through which the sensor collects parameter information such as acceleration and position of each part of the limb to realize the analysis and recognition of the posture. Noncontact recognition technology is usually based on vision-based human gesture recognition, that is, extracting and recognizing human features through video surveillance, which is a hot topic in computer vision.

2.1.1. Contact Recognition Technology. Contact human body gesture recognition technology collects motion parameters through sensors worn on the human body and analyses the parameter information to realize gesture recognition. Contact recognition technology can accurately capture the changes in posture by allowing users to wear data collection equipment, and the posture recognition rate after computer analysis is quite high. However, in practical applications, it will cause inconvenience to users and does not meet the requirements of human-computer interaction. At the same time, such equipment is expensive and difficult to be popularized in real life.

2.1.2. Noncontact Recognition Technology. Vision-based noncontact human posture recognition technology is the hotspot in the field of human-computer interaction. It acquires image information through video capture devices such as cameras, and after computer processing and analysis, specific representative features are extracted and classified to achieve posture recognition.

2.2. Algorithm Classification. According to different classification standards, it can be divided into a variety of human body gesture recognition algorithms. From the perspective of implementation methods, it is usually divided into three types:

(1) Three-dimensional model reconstruction method, which extracts three-dimensional features from valid samples to construct a three-dimensional model

(2) The human body appearance model method, which establishes a two-dimensional model by acquiring the shape characteristics of the human body and uses the model matching method to complete the recognition

(3) The motion model method is classified according to motion characteristics

From the perspective of pattern recognition, gesture recognition is a problem of classifying time-varying features, that is, the process of matching the test sequence with the present sequence according to the obtained feature information. From the perspective of matching methods, there are usually two types.

2.2.1. Template Matching Method. The template matching method regards the human body motion posture as a combination of a series of discrete static images in a certain period of time. By extracting the static features of each discrete image from them and matching them with the trained posture model, the highest matching degree of motion posture is the best result.

2.2.2. State Space Method. The state space method is to set each key static posture to a specific state. And, all dynamic postures are a traversal process that connects these specific states with the best probability relationship, and the best probability at this time is set as the classification standard of this posture. However, the state space method finds the global optimal solution through complex iterative operations, so it is difficult to apply it in practical work.

3. Improved Generative Adversarial Network Algorithm

3.1. Overall Framework. As shown in Figure 1, the main steps of human posture recognition based on depth information are divided into three parts: first, we preprocess the image data collected by the depth information sensing device and then extract the corresponding human posture image features, which is the ROI area, according to the different targets and finally use appropriate the classification algorithm which performs posture classification and recognition. After the training and classification process, you can view the system's classification and recognition results of a new input image, select a picture from the test sample set, and input it into the network. According to the results of the softmax classifier, the label with the highest probability is the current recognition result. The final recognition result will appear in the following two situations: when the system predicted classification does not match the expected label, the classification error is displayed; when the test result matches the expected pose label, the classification is displayed correctly.

3.2. Deep Convolutional Hourglass Network. The basic unit of the deep convolutional network is the residual network, as shown in Figure 2. It mainly includes three convolutional layers. The size of their convolution kernels is different. The size of the first layer is 1×1, and the size of the second layer is 3×3, and the size of the third layer is 1×1. Before passing through each convolutional layer, a Batch Normalization layer and a ReLU activation layer will be passed. The network extracts high-level features through the convolutional layer of the main path. The branch is a jump layer, which is composed of a convolutional layer with a core size of 1×1. The main function is to retain the features of the original layer, increase the nonlinearity of the model, and reduce the amount of calculation. In the residual network, only the number of channels of the image is changed, and the step size of all convolutional layers is 1. If the channel of the input image is M and the channel of the output image is N, the number of the first and second convolutional layer kernels of the main path is $N/2$ and the number of the third convolutional layer kernel is N.

Based on the residual network, a first-order hourglass network is constructed following its structure. Both the main loop and the branch include several residual networks. Among them, the branch is extracted at the original size through the jump layer, which can well retain the spatial information of each joint on the picture; the main road first uses the maximum pooling layer to change the size to half of the original size and then performs feature extraction. Through this method, the network extracts the joint point features at different resolutions and finally restores the size to the original size through nearest neighbour interpolation and adds it to the branch output. Defining the feature of the image output through the first-order hourglass network as $F1\ (I\ (x))$, the following equation is obtained:

$$F_1(I(x)) = C(x) + \operatorname{deconv}(C(\operatorname{Maxpool}(I(x)))). \quad (1)$$

Among them, $C\ (x)$ is the output feature of the residual network.

By replacing the residual network inside the dotted line in the first-order hourglass network with the first-order hourglass network, a second-order hourglass network can be obtained. In the same way, by replacing it in the third-order hourglass network, you can get a fourth-order hourglass network. In the node detection network, a stacked four-stage hourglass network is used as the detection network. Among them, each branch is branched before passing the maximum pooling layer to preserve the spatial information between the joints on the picture. Three residual modules are used for feature extraction after each down-sampling. Therefore, the 4-order hourglass network can extract the joint point features on the original size, 1/2, 1/4, and 1/8 size. After each feature extraction, the image is restored to its original size through up-sampling. After adding the original size features, a residual network is used for feature extraction. Therefore, the network does not change the size of the image, only the number of feature channels. Defining the output feature of the image through the fourth-order hourglass network as $F4\ (I\ (x))$, it can be briefly described as follows:

$$F_4(I(x)) = F_3\left(F_2\left(F_1(I(x))\right)\right). \quad (2)$$

3.3. Improved Model. Since the detection of the original size of the image requires a lot of time, the pixels are first reduced to 64×64 through the convolutional layer and the pooling layer, and the residual network is used for feature extraction in the middle. The structure is as follows:

(1) Turn the 256×256 pixel RGB image through 64 7×7 convolution kernels with a step size of 2 to 128×128 pixels, and the number of I channels becomes 64.

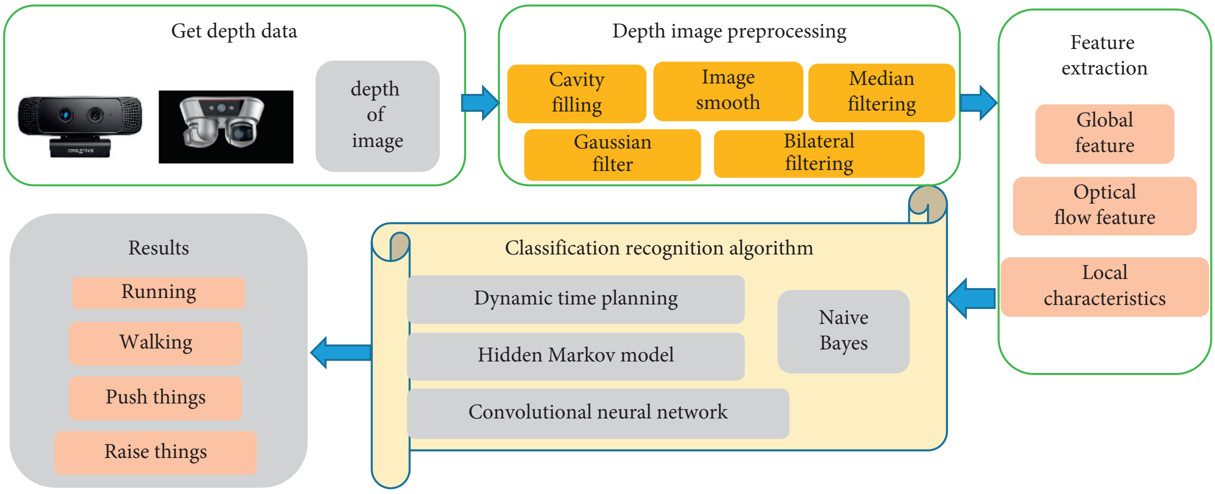

Figure 1: Flow chart of human body gesture recognition based on depth information.

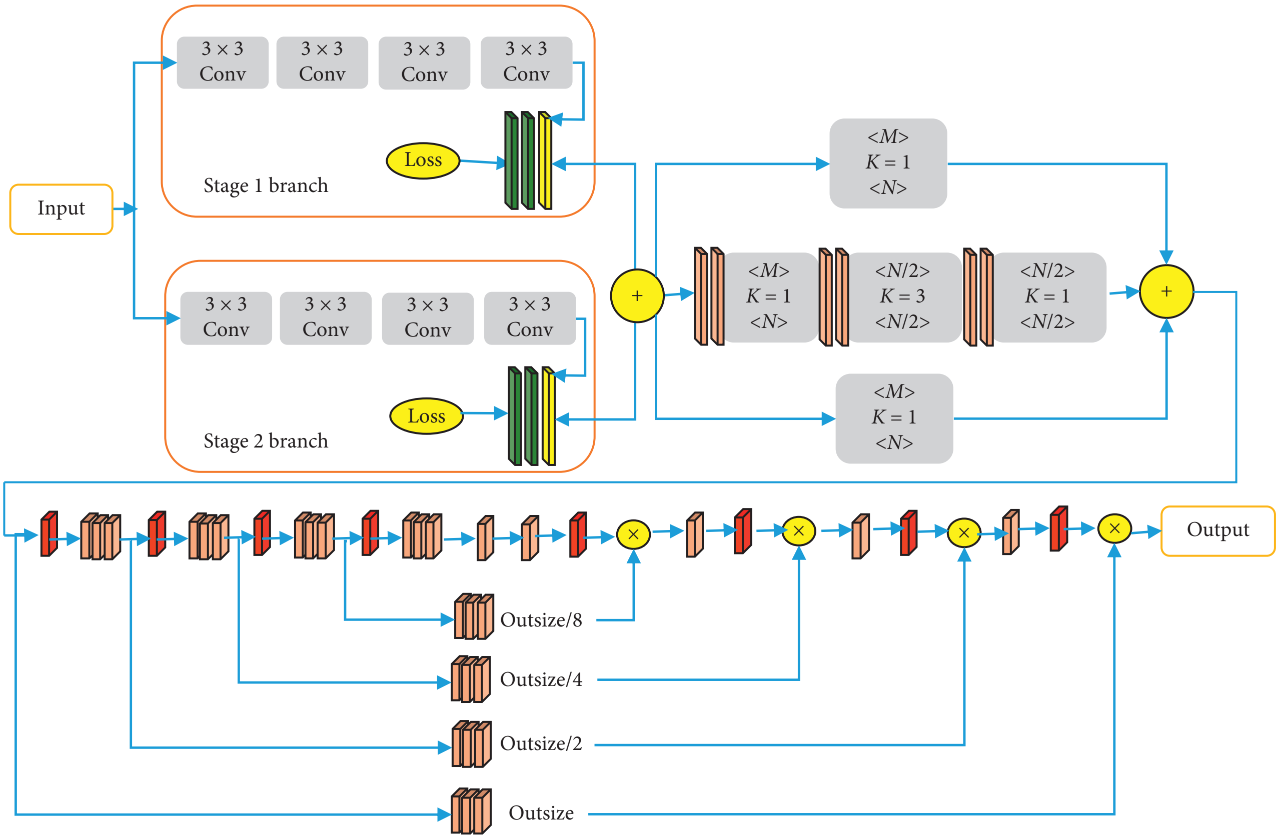

Figure 2: Residual network based on the hourglass model.

$$O_k = \mathrm{Relu}\left(w * I(k) + b_k\right). \tag{3}$$

(2) Through the residual network, the input channel is 64 and the output channel is 128.

$$O_m = F\left(O_k\right). \tag{4}$$

(3) Carry out maximum pooling and further reduce the data pixel 128 × 128 to 64 × 64 without changing the number of channels.

$$O_p = \mathrm{Maxpool}\left(O_m\right). \tag{5}$$

(4) In the same way, three consecutive residual networks are used after down-sampling. The input and output channels of the first two networks are both 128, and the last output channel is 256.

$$O_f = \mathrm{Maxpool}\left(F\left(\mathrm{Relu}\left(w * I(k) + b_k\right)\right)\right). \tag{6}$$

The model proposed in this paper is DCGAN, which includes two parts: generator and discriminator. As shown in Figure 3, Stack 1, Stack 2, and Stack 3, respectively, represent different motion postures. In addition, the basic structure of the generator and discriminator is based on the hourglass network, which is an encoder and decoder network with jump connections and up-sampling layers, which can adapt to images of different sizes while capturing the diversity of human body poses. The input and output of the hourglass network are preprocessed and postprocessed using residuals and linear blocks to extract features and estimate the position of body parts to obtain a valuable confidence score map.

The position of the joints of the limbs of the human body belongs to the articulated structure, which leads to obvious deformation of the limbs when performing movements, and in most cases, the degree of deformation of the joints on the trunk is relatively small. Therefore, it is more challenging to estimate the posture of the parts of the human body with large deformation, and a large number of different training samples are required as support. In order to solve these problems, this research optimizes the structure of generator and discriminator networks. The new design uses stacked hourglass networks, which can be connected to build the spatial hierarchy of human body parts. In addition, this study also introduced a new loss function term to regulate the relationship between parents and children. The hierarchical adversarial network design and new hierarchical perceptual loss help to accurately estimate the position of various parts of the body, especially for those body parts that are highly deformed or highly occluded.

3.3.1. Hierarchical Generation Network. The generator maps the input image to the confidence maps P and C of the parent and child, respectively. The mapping can be performed through the learning function as follows:

$$\begin{aligned} P_k &= G_k\left(I(x), L\left(P_{k-1}, C_{k-1}\right)\right), \\ C_k &= G_k\left(I(x), L\left(P_k\right), L\left(P_{k-1}, C_{k-1}\right)\right). \end{aligned} \tag{7}$$

Then, connect the confidence map of the parent part and the child part to the output of each stack generator. The predicted pose can be inferred using the following equation:

$$Z = \operatorname{softmax}\left(\sum_{i=1}^{n}\left(P_k, C_k\right)\right). \tag{8}$$

Estimate the predicted pose Z by summing the confidence maps of all sequential connection stacks and using softmax on the confidence scores of the connection maps. This allows the entire process to be trained end-to-end: the generator network itself has no adversarial branch and can only be trained by minimizing the following loss function:

$$\text{Loss}_G = \min\left(\sum_{i=1}\sum_{j=1}\left|P_i^j - C_i^j\right|^2\right). \tag{9}$$

3.3.2. Hierarchical Discrimination Network. Training the discriminator is a very important step because it is more inclined to reconstruct the discriminator from the generated pose than from the generated pose, which makes the discriminator unable to distinguish whether the generated pose is true or false. In order to solve this problem, this paper uses a balance strategy between reconstructing the real pose and generating the pose. Use the following loss function to train the discriminator:

$$\text{Loss}_D = \min\left(\alpha * \text{HD} + \beta * \text{HR}\right). \tag{10}$$

Among them, α and β are balance terms, and HR and HD are the loss terms of the real pose and the generated pose, respectively.

4. Results and Discussion

4.1. Experimental Data and Experimental Environment. In this chapter, in order to verify the performance of the proposed network, three challenging human pose estimation data sets will be used to conduct experiments and get the corresponding results.

(1) The LSP data set and its extended version here contain 12,000 images, of which 11,000 images will be used for training and the other 1,000 will be used for testing.

(2) MPII data set contains 25,000 images and 40,000 images of human actions.

(3) The LIP data set contains 50,000 images, of which 16 key annotations are used for human pose estimation. The dataset is collected from real scenes with various poses and views and also contains many images with large occlusion and low resolution.

(4) Create your own data set. This article collects 200 images of 5 people in 10 postures to build a posture training database. The 10 postures are standing, squatting, raising hands, walking, folding hands, crossing hands, bending over, punching, raising legs, and bowing and groups the collected images from pose 1 to pose 10.

Under laboratory conditions, the hardware equipment needed for this paper are as follows:

(1) One computer with win10 system and Linux system is needed. Due to the special requirements of the Kinect SDK, the computer needs to be a 64-bit system, the computer CPU is Intel (R) Core (TM) i7-2640M, the CPU frequency is 2.80 GHz, and the installed memory is 8.00 GB. The quality of the computer configuration will affect the subsequent classifier training and classification efficiency. The experiments in this paper are all carried out on this computer.

(2) Kinect for the windows camera is needed. The camera's parameters are shown in Table 1.

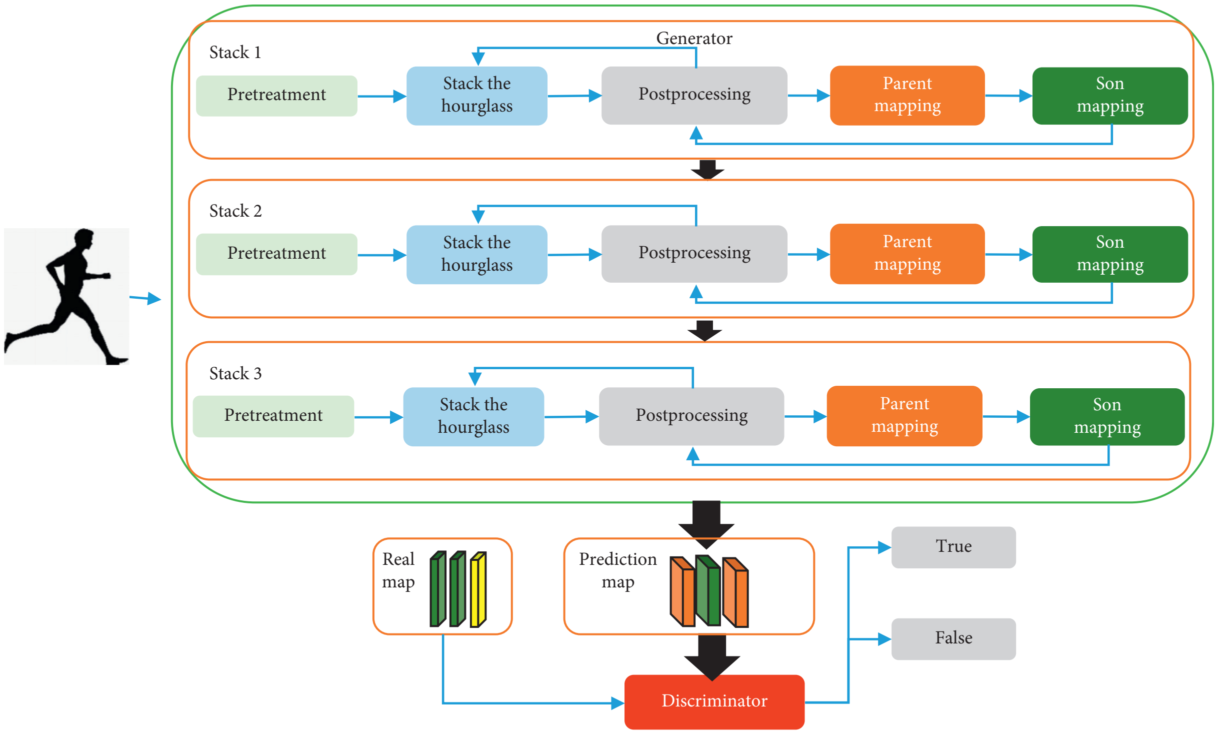

Figure 3: DCGAN framework.

Table 1: Kinect camera parameter.

Induction project	Effective coverage
Color and depth	1.2–3.6 meters
Skeleton tracking	1.2–3.6 meters
Vision	57 degrees horizontally, 43 degrees vertically
Induction project	Effective coverage
Base motor rotation	28 degrees left and right
Frames per second	30 FPS
Depth resolution	QVGA (640 × 480)
Color resolution	VGA (640 × 480)
Sound format	16 KHz, 16-bit mono pulse code modulation (PCM)
Voice input	Four-microphone array, 24-bit analog-to-digital conversion (ADC), noise cancellation

(3) 220V AC power supply is needed. During the experiment, the image acquisition part uses Visual Studio2015 as the development environment, uses KinectSDK for Windows to operate the data stream of the camera, and the Opencv function library organizes the collected data stream into the required human posture image and stores it on the mobile hard disk. The subsequent CNN training and gesture recognition algorithm are completed in Caffe.

4.2. Network Training. First, initialize the network and input each set of pose samples in the training set into the previously constructed CNN system and train one pose in one round. As shown in Figure 4, with the increase in training samples, the loss function value of the network continues to decrease, and the accuracy of verification continues to increase. And, when the training sample reaches 800, it starts to converge and tends to stabilize so as to determine the parameter value of the network. It can be seen that 1000 training samples can fully meet the training needs of the network.

4.3. Algorithm Recognition Performance Analysis

4.3.1. LSP Data Set. The result is shown in Figure 5. The percentage of correct key points for the model proposed in this paper can reach 94.2%. This result is better than the results in [21, 23]. These two methods are also considered to the method proposed in this paper and is the closest comparable method. Although the literature [22] and the literature [23] can estimate the structure information of the human body more accurately, the model proposed in this paper can estimate the pose on challenging body parts such as elbows, wrists, knees, and ankles. This shows that the

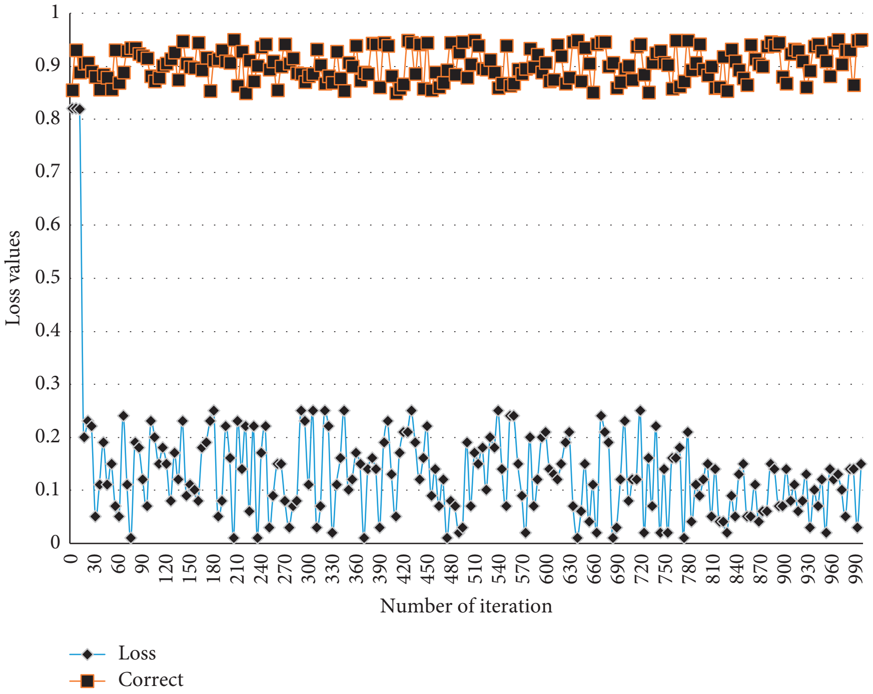

Figure 4: Accuracy change and loss change of training sample increase.

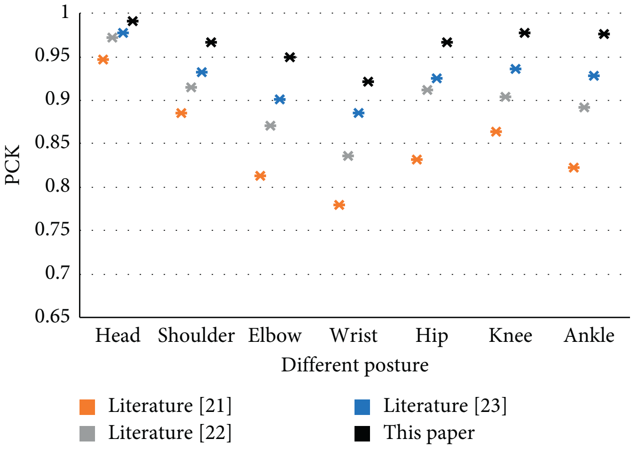

Figure 5: The percentage of correct key points in the LSP data set (PCK).

proposed network can not only achieve the effectiveness of estimating human posture but also consider the hierarchical structure of various body parts.

4.3.2. LIP Data. Figure 6 shows the PCKh results of this method and compares them with the results of the previous method. PCKh is the same as ordinary PCK, and the difference is that the tolerance is calculated based on the head instead of the torso. The model proposed in this paper is better than other results that use training these data sets for pose estimation. Similarly, the model in this paper focuses on the hierarchical relationship of the human body structure, which can be better detected in key parts that are prone to large deformations. The accuracy of the comparison between the wrist part and the method in [22] can be explained this point. Hybrid pose machine [30] and pyramid stream network [31] are generally considered to be classic methods for training on LIP data sets. This paper compares these two methods, and the results show that the method in this paper is better. This also confirms the effectiveness of the proposed method and the importance of encoding the hierarchical structure into the model.

4.3.3. MPII Data Set. The MPII data set represents the most challenging benchmark data set in human pose estimation because it contains many changes in various poses of people performing different activities. In Figure 7, the PCKh results are shown in the test set using the MPII data set, and all results have a tolerance of 0.5. Compared with the methods in the literature [22, 23] that do not consider the hierarchical

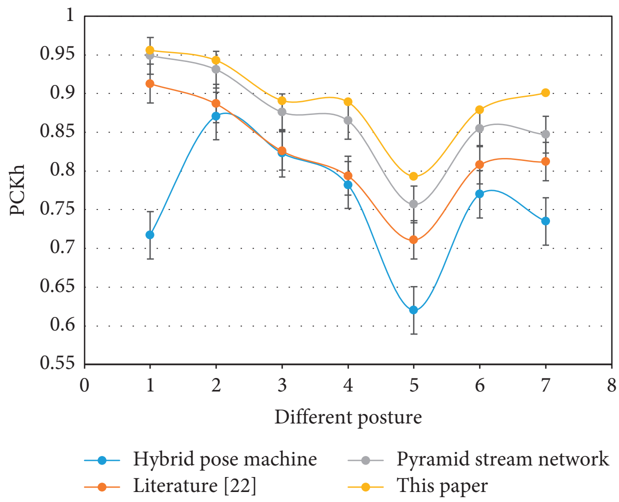

Figure 6: The percentage of correct key points in the LIP data set (PCKh).

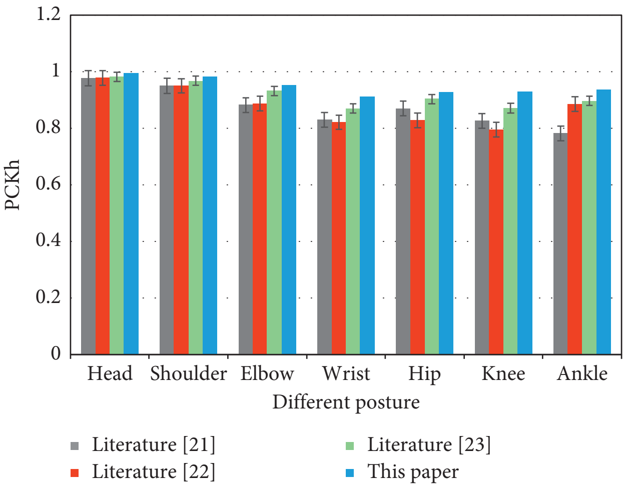

Figure 7: The percentage of correct key points in the MPII dataset (PCKh).

structure, this method can show better results for most body parts estimation.

The PCKh value of this method is better than the results of other methods for the average value of all parts and the value of the wrist and ankle parts. This once again proves the robustness of the proposed method in estimating the poses of difficult body parts.

Since the MPII data set is the most challenging in human pose estimation, this article will conduct comparative experiments with different tolerances for this data set. Figure 8 shows the average PCKh value on the MPII data set. It can be seen from the curve that the method in this paper is better than existing methods under low tolerance.

4.3.4. Self-Created Data Set. Through the training and recognition of the neural network, the experimental results obtained are shown in Figure 9. It can be seen from Figure 9 that the recognition accuracy rate has reached more than 90%, and the expected accuracy rate goal has been achieved.

Because different users have individual differences, such as tall, short, fat, and thin, there will be user independence problems in the human body gesture recognition process, and the human body gesture features extracted by CNN in this paper are invariant to the user's posture changes. Therefore, in order to verify the versatility of the recognition method in this paper, 450 pictures of two fatter and smaller users that are not included in the human body pose database are selected to verify the recognition effect of this paper, and each type of sample is 150. The experimental results are shown in Table 2.

When the user's posture changes, the classification accuracy of the recognition method in this paper is almost unaffected. This shows that the human posture feature extraction method in this paper can well characterize the characteristics of human posture and to a certain extent solves the user independence in the process of human posture recognition.

In the past, human body gesture recognition is often limited by lighting conditions, and the recognition effect is often limited in an environment with too dark or too bright light. In order to verify the recognition effect of the recognition algorithm in this paper in a universal environment, this paper collected 360 sample images of the human body pose of the same person under different lighting and background conditions. Among them, there are 120 different kinds of human body postures. It was identified, and the results are shown in Table 3.

The experimental results show that when environmental factors such as light and background change, the classification accuracy of the recognition method in this paper is almost unaffected, and the overall average accuracy rate reaches 0.965, which shows that the recognition algorithm in this paper has universal applicability to the environment.

Collect 300 images containing subjects under joint occlusion and joint nonocclusion mainly to calculate the deviation distance between the actual coordinates of the shoulder, elbow, wrist, and knee joints and their theoretical coordinates and the maximum error of the human body gesture recognition in this article 3 cm for comparison. After calculation, when the joint is not occluded, the joint recognition accuracy of this paper is 0.82, as shown in Figure 10. When the joints (wrists and knees) are blocked, the joint recognition accuracy can also reach 0.98.

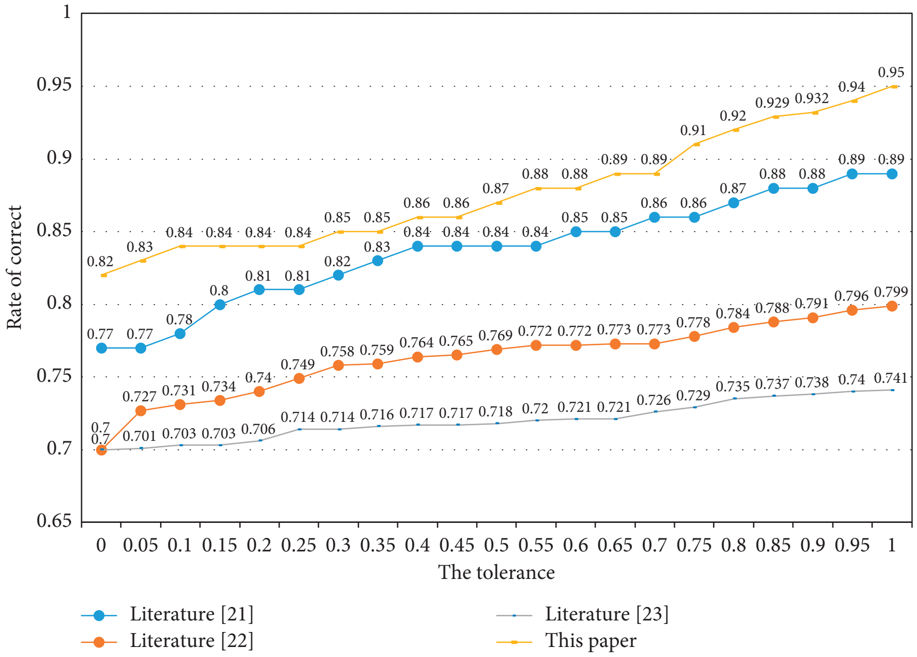

Figure 8: Average PCKh curve under different tolerances.

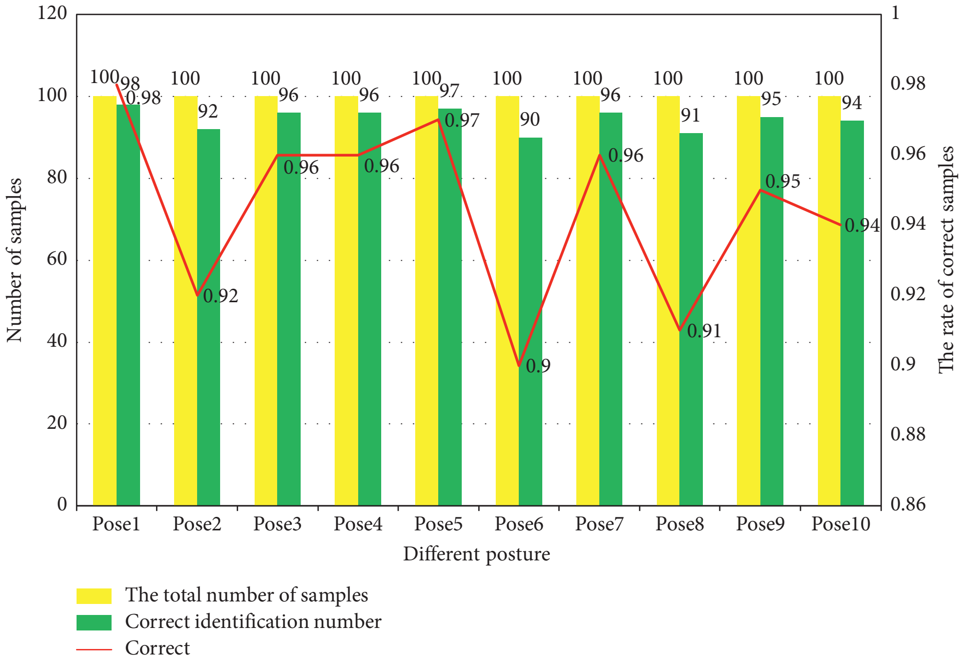

Figure 9: Analysis of recognition results.

Table 2: Kinect user independence recognition accuracy rate.

	Standing	Squats	Take your hands
The total number of samples	150	147	0.980
Correct identification number	150	136	0.906
Correct	150	141	0.940

Table 3: Kinect universal environment recognition accuracy rate.

	Standing	Squats	Take your hands
The total number of samples	120	197	0.985
Correct identification number	120	186	0.930
Correct	120	196	0.980

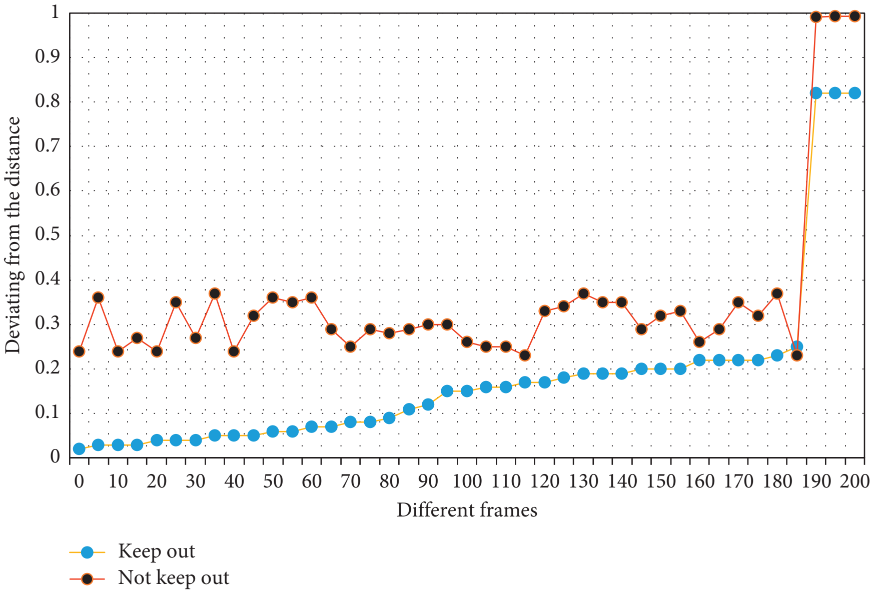

Figure 10: The distance between the actual and theoretical coordinates of the occluded and un-occluded joints.

5. Conclusion

This paper proposes a generated confrontation network based on stacked hourglasses, which can implicitly infer the structure and hierarchy of various body parts. The network proposed in this article shows the ability to learn to estimate the spatial relationship of body parts. It is trained end-to-end and tested on three different benchmark data sets. The network has shown the ability to estimate body parts, which are largely deformed and highly occluded. The method proposed in this paper can overcome these obstacles and achieve better performance. It also obtains the latest experimental results on a variety of data sets and compares them with the results on other related data sets. In addition, the network proposed in this paper can also be extended to encode other associated features between adjacent body parts. Moreover, the network is also suitable for processing multiple hierarchical levels between body parts. Currently our algorithm is more sensitive to light, which is also the limitation of our algorithm. Our next work is to solve the problem that the algorithm is sensitive to light.

Data Availability

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

Conflicts of Interest

The authors declare that they have no known conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant no. U1804152.

References

[1] D. Jiang, G. Li, Y. Sun, J. Kong, and B. Tao, “Gesture recognition based on skeletonization algorithm and CNN with ASL database,” *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 29953–29970, 2019.

[2] Q. De Smedt, H. Wannous, and J.-P. Vandeborre, “Heterogeneous hand gesture recognition using 3D dynamic skeletal data,” *Computer Vision and Image Understanding*, vol. 181, pp. 60–72, 2019.

[3] C. Wang, Z. Zhang, and Z. Xi, “A human body based on sift-neural network algorithm attitude recognition method,” *Journal of Medical Imaging and Health Informatics*, vol. 10, no. 1, pp. 129–133, 2020.

[4] X. Liu, K. N. Khan, Q. Farooq, Y. Hao, and M. S. Arshad, “Obstacle avoidance through gesture recognition: business

advancement potential in robot navigation socio-technology," *Robotica*, vol. 37, no. 10, pp. 1663–1676, 2019.

[5] X. Dong, Y. Lei, T. Wang et al., "Automatic multiorgan segmentation in thoraxCTimages using U-net-GAN," *Medical Physics*, vol. 46, no. 5, pp. 2157–2168, 2019.

[6] B. Li, L. Liu, M. Shen, Y. Sun, and M. Lu, "Group-housed pig detection in video surveillance of overhead views using multi-feature template matching," *Biosystems Engineering*, vol. 181, pp. 28–39, 2019.

[7] Z. Zhang, H. Zhang, and T. Liu, "Study on body temperature detection of pig based on infrared technology: a review," *Artificial Intelligence in Agriculture*, vol. 1, pp. 14–26, 2019.

[8] J. Cui, C. Min, and D. Feng, "Research on pose estimation for stereo vision measurement system by an improved method: uncertainty weighted stereopsis pose solution method based on projection vector," *Optics Express*, vol. 28, no. 4, pp. 5470–5491, 2020.

[9] Y. Zhang and X. Lu, "Measurement method for human body anteflexion angle based on image processing," *International Journal of Imaging Systems and Technology*, vol. 29, no. 4, pp. 518–530, 2019.

[10] L. Li, X. Chen, J. Wu, S. Wang, and G. Shi, "No-reference quality index of depth images based on statistics of edge profiles for view synthesis," *Information Sciences*, vol. 516, pp. 205–219, 2020.

[11] W. Zhang, D. Kong, S. Wang, and Z. Wang, "3D human pose estimation from range images with depth difference and geodesic distance," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 272–282, 2019.

[12] P. Duan, T. Wang, M. Cui, H. Sang, and Q. Sun, "Multi-person pose estimation based on a deep convolutional neural network," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 245–252, 2019.

[13] A. Nadeem, A. Jalal, and K. Kim, "Accurate physical activity recognition using multidimensional features and Markov model for smart health fitness," *Symmetry*, vol. 12, no. 11, p. 1766, 2020.

[14] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K.-K. R. Choo, "Imaging and fusing time series for wearable sensor-based human activity recognition," *Information Fusion*, vol. 53, pp. 80–87, 2020.

[15] Z. Xu, W. Lu, Q. Zhang, Y. Yeung, and X. Chen, "Gait recognition based on capsule network," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 159–167, 2019.

[16] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks," *IEEE Access*, vol. 7, pp. 38044–38054, 2019.

[17] A. Jalal, S. Kamal, and C. A. Azurdia-Meza, "Depth maps-based human segmentation and action recognition using full-body plus body color cues via recognizer engine," *Journal of Electrical Engineering & Technology*, vol. 14, no. 1, pp. 455–461, 2019.

[18] Y. Alshawabkeh, "Linear feature extraction from point cloud using color information," *Heritage Science*, vol. 8, no. 1, pp. 1–13, 2020.

[19] B. Luo, Y. Sun, G. Li, D. Chen, and Z. Ju, "Decomposition algorithm for depth image of human health posture based on brain health," *Neural Computing and Applications*, vol. 32, no. 10, pp. 6327–6342, 2020.

[20] M. A. Jaffar, M. S. Zia, M. Hussain, A. B. Siddiqui, S. Akram, and U. Jamil, "An ensemble shape gradient features descriptor based nodule detection paradigm: a novel model to augment complex diagnostic decisions assistance," *Multimedia Tools and Applications*, vol. 79, no. 13-14, pp. 8649–8675, 2020.

[21] M. S. Alzahrani, S. K. Jarraya, H. Ben-Abdallah, and M. S. Ali, "Comprehensive evaluation of skeleton features-based fall detection from Microsoft Kinect v2," *Signal, Image and Video Processing*, vol. 13, no. 7, pp. 1431–1439, 2019.

[22] S. Ghazal, U. S. Khan, M. Mubasher Saleem, N. Rashid, and J. Iqbal, "Human activity recognition using 2D skeleton data and supervised machine learning," *IET Image Processing*, vol. 13, no. 13, pp. 2572–2578, 2019.

[23] A. Franco, A. Magnani, and D. Maio, "A multimodal approach for human activity recognition based on skeleton and RGB data," *Pattern Recognition Letters*, vol. 131, pp. 293–299, 2020.

[24] W. Ren, O. Ma, H. Ji, and X. Liu, "Human posture recognition using a hybrid of fuzzy logic and machine learning approaches," *IEEE Access*, vol. 8, pp. 135628–135639, 2020.

[25] J.-i. Lee, J. B. Han, J.-B. Han et al., "A study on design of posture transition filter for 3D human posture estimation and refinement on robotic bed," *Journal of Korea Robotics Society*, vol. 15, no. 3, pp. 269–276, 2020.

[26] M. Y. A. Khanian, S. M. R. H. Golpayegni, and M. Rostami, "A new multi-attractor model for the human posture stability system aimed to follow self-organized dynamics," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 162–172, 2020.

[27] Q. Ke, J. Zhang, W. Wei et al., "A neuro-heuristic approach for recognition of lung diseases from X-ray images," *Expert Systems with Applications*, vol. 126, pp. 218–232, 2019.

[28] L. Zhang, P. Shen, X. Peng et al., "Simultaneous enhancement and noise reduction of a single low-light image," *IET Image Processing*, vol. 10, no. 11, pp. 840–847, 2016.

[29] B. Zhou, X. Duan, D. Ye et al., "Multi-level features extraction for discontinuous target tracking in remote sensing image monitoring," *Sensors*, vol. 19, no. 22, p. 4855, 2019.

[30] Y. Shen and K. Khorasani, "Hybrid multi-mode machine learning-based fault diagnosis strategies with application to aircraft gas turbine engines," *Neural Networks*, vol. 130, pp. 126–142, 2020.

[31] X. Yang, Y. Lei, T. Wang et al., "PET attenuation correction using MRI-aided two-stream Pyramid attention network," *Journal of Nuclear Medicine*, vol. 61, no. supplement 1, p. 110, 2020.