3D Dynamic Hand Gesture Recognition with Fused RGB and Depth Images

Qingshan Ye, Yong Bai, Lu Chen and Wenjie Jiang

Hainan University, No. 58, Renmin Avenue, Meilan District, Haikou,570208, China.

Abstract

With the advancing of dynamic gesture recognition technology, it is widely used in various interaction scenarios nowadays. However, the three-dimensional dynamic gesture recognition method is easily disturbed by the external environment, such as illumination, background and shadow. To address these problems, we propose a 3D Dynamic Gesture Recognition network model, which use both CNN and LSTM networks and can fuse RGB and depth image information. We conduct experiments with Intel RealSense D415 depth camera and self-built gesture dataset, the results demonstrate that the recognition accuracy of the 3D-DGR model is 99.23%, which is 1.52% higher than the model using only RGB images.

Keywords

Dynamic gesture recognition, depth camera, human-computer interaction

1. Introduction

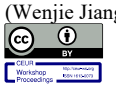
Human-Computer Interaction refers to the process of communication between human and computer with the help of some methods. In recent years, with the development of science, technology and the popularization of smart devices, Human-Computer Interaction technology has become more and more widely used in people's daily life. Gestures are one of the common expressions in people's life, and with the continuous development of society, gestures have gradually evolved to have certain meanings and become a powerful way and means of expressing one's emotions in addition to language, hence gesture-based interaction has become an important element of human-computer interaction [1]. In the early days, Human-Computer Interaction has carried out[1] through wired devices, such as mouse and keyboard, which became the main Human-Computer Interaction mode and has been continued until now, but such input method is restricted by hardware devices and cannot be used at will. As the technology of Human-Computer Interaction becomes more mature, the human-computer interaction mode which is more in line with the interactive experience starts to appear in people's vision. Gesture recognition refers to the computer's technical methods to recognize human gestures and discuss their important meanings. Compared with traditional interaction by hardware devices such as mouse and keyboard, it can realize communication between human and computer without contacting the machine, which greatly improves people's interactive experience. To quickly identify the gesture, early gesture recognition is generally based on wearable devices to obtain finger and other motion data. The advantage of this approach is that it is real-time, has high recognition accuracy and is not affected by external factors such as lighting, color and camera pixels [2], but its disadvantage is that the device is expensive, restricts the movement of the operator and has a poor user experience.

The later development of gesture recognition technology through ordinary camera-based gestures is a process from two-dimensional (2D) to three-dimensional (3D) [3]. Gesture recognition based on traditional 2D images, also known as 2D gesture recognition. It recognizes the simplest type of gestures,

ICBASE2022@3rd International Conference on Big Data & Artificial Intelligence & Software Engineering, October 21-23, 2022, Guangzhou, China

EMAIL: 20085400210192@hainanu.edu.cn (Qingshan Ye); bai@hainanu.edu.cn (Yong Bai); 20081000210002@hainanu.edu.cn (Lu Chen); 20085400210135@hainanu.edu.cn (Wenjie Jiang)

ORCID:0000-0002-8283-7554(Qingshan Ye); 0000-0002-2506-5981 (Yong Bai); 0000-0001-9152-4237 (Lu Chen); 0000-0002-6236-8623 (Wenjie Jiang)



normally specified static gestures, such as a clenched fist or an open five-finger gesture [4]. This technique can only recognize the "state" of a gesture, but not the continuous change of a gesture [5]. Moreover, this visual technique is usually affected by factors such as light and skin color, resulting in the inability to recognize the user's intention to use the gesture or the low accuracy of the recognition. 3D gesture recognition has an additional temporal dimension than two-dimensional gesture recognition, allowing the recognition of dynamic gestures and the perception of continuous changes in gestures. This dynamic gesture recognition can recognize a wider range and achieve more functions than the 2D gesture recognition [6][7]. However, it is still susceptible to factors such as light, skin color, and background, which leads to a low recognition accuracy. With the emergence of depth cameras, the depth information obtained from depth sensors can better exclude problems such as background lighting and color information sensitivity, but the process of recognizing gestures still suffers from interference such as environmental objects. Therefore, gesture recognition with fused RGB image and depth image has the potential to resolve the above problems. In this paper, we propose a 3D dynamic gesture recognition network model that fuses RGB images and depth images in CNN and LSTM networks, and apply the model to the human-computer interaction for web page navigation using Intel RealSense depth camera.

2. Related Work

The main traditional machine learning methods are Dynamic Temporal Regularization (DTW) [8], Hidden Markov model (HMM) [9], Conditional Random Field (CRF) [10], and Random Forest (RF) methods [11]. Major deep learning methods are based on CNN and LSTM networks [12]. Traditional machine learning methods are less demanding in terms of training data and computational power, but the accuracy is usually lower than the deep learning-based methods. DTW is a template matching algorithm, which is relatively simple to implement and does not need training, but requires high-precision templates for matching [13]. HMM and CRF methods are both probabilistic model-based algorithms, and both of them can extract dynamic temporal [14]. The RF algorithm, as a common machine learning algorithm, mainly uses an integrated tree classifier [15]. With the development of deep learning in recent years, object detection algorithms such as Faster RCNN and SSD have been increasingly applied to gesture recognition [16][17], which has the advantages of higher accuracy and better robustness. Then some researchers proposed hand key point detection method [18], which is a method with greater development potential because the hand pose is not affected by the background information and can better focus on the position and motion information of the hand compared with the RGB image based gesture recognition method, but the algorithm model is complex and requires high computational power. For the gesture recognition problem, both traditional machine learning-based methods and deep learning-based gesture recognition methods generally need to extract the location of the hand in the video first, which is also called hand detection. Traditional hand detection has methods based on hand color and hand motion information [19]. Hand color-based methods use the difference between hand color and background color information for hand segmentation, but this method is sensitive to background lighting, color information. Hand motion information-based methods use hand motion information relative to the background for gesture segmentation, this method requires the background information to be approximately constant, and has a less robust. 3D gesture recognition can perceive the continuous change of gestures, and can recognize more gestures containing semantic information. However, it is still vulnerable to the effects of illumination and background information.

With the popularity of depth sensors, the depth information obtained from depth sensors, some researchers proposed to reconstruct the 3D information of gestures using only depth map information for recognition [20], but the algorithm model is complex and cannot learn the representational information of gestures. In order to further resolve the above problems, we propose an 3D-DGR, which is a 3D dynamic gesture recognition network model that can fuse RGB image and depth image information using CNN and LSTM. In such a model, the RGB images and depth images captured by Intel RealSense camera are fed into CNN and LSTM to extract spatial and temporal features and then fused for gestion classification.

3. Methods

Aiming at the problems of environmental objects and insufficient utilization of depth information in dynamic gesture recognition, we propose a 3D dynamic gesture recognition network model 3D-DGR that integrates RGB and depth images information. The overall architecture of our 3D-DGR is shown in Figure 1. In addition, the connection between the fully connection layers (FC3 and FC4) and LSTM is illustrated in Figure 2. The operations of CNN and LSTM are represented in the legends of arrows with different colors.

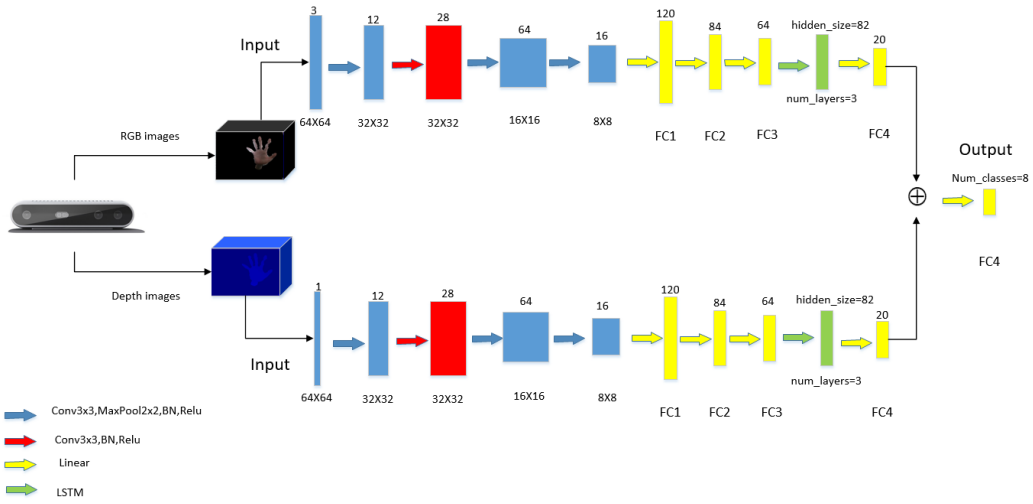


Figure 1: 3D-DGR Network Architecture

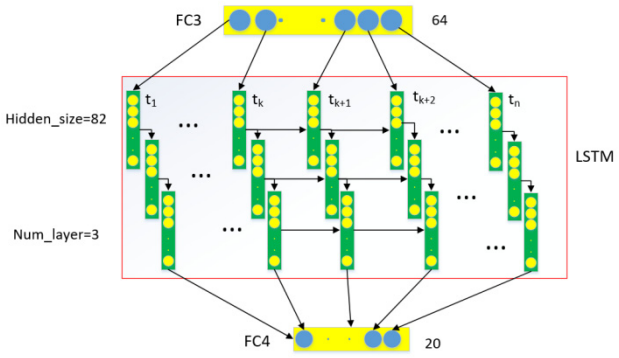


Figure 2: The connection between the fully connection layers (FC3 and FC4) and LSTM

In our experiment setup, the Intel RealSense D415 depth camera acquires color and depth image data simultaneously. Figure 2 shows the RGB image and depth image acquired by using the D415 camera. It can be seen that the objects have different colors when they are not at the same distance from the camera. The closer the object to the camera, the darker the blue color.

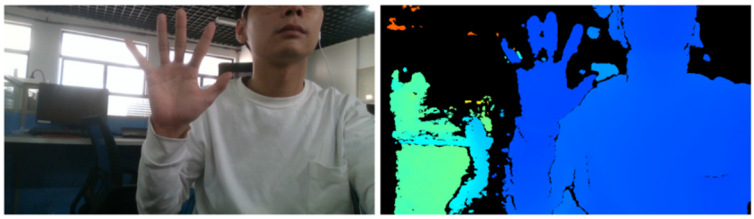


Figure 3: D415 depth camera captured images. (a) RGB image. (b)Depth image

In Figure 3, it can be seen that there are still environmental interference factors. To reduce the environmental interference and use the information of the depth image data, a distance threshold can be set so that if the object is higher than this distance threshold, that is, farther away from the camera, it will not appear in the recognition area, which greatly reduces the interference characteristics of the environment. Hence the model pays more attention to the gestures that appear in the recognition area, and it can segment gesture images better and has enhanced robustness in the usage. The hand gesture images after depth thresholding segmentation is shown in Figure 4.

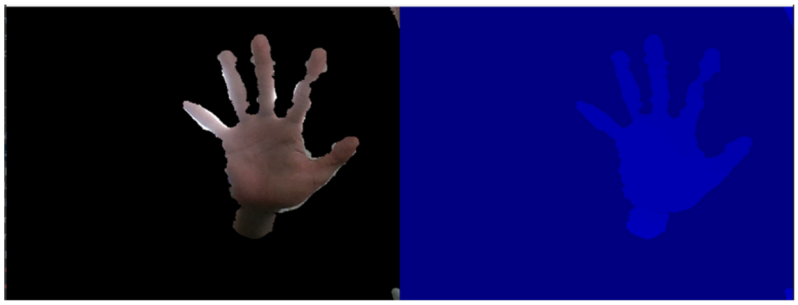


Figure 4: Hand gesture images after depth thresholding segmentation. (a) RGB image. (b) Depth image

Next, we describe the detailed process in our proposed model. First, the input image is pre-processed with a depth threshold to reduce the interference from the surrounding environment, so that the network can focus more on the gesture action itself. In the experiment, we set the depth threshold value to 0.5m. The target is automatically shielded from the depth camera greater than 0.5m, which is shown as black in the RGB image. Then the pre-processed RGB image and depth image are fed into the network model. The 3D-DGR model is divided into two main branches, the upper branch is responsible for processing RGB images and the lower branch is responsible for processing depth images. Each branch consists of two main parts, a CNN module is responsible for feature extraction in spatial domain and a LSTM module is responsible for processing the extracted features in time domain. In CNN module, 4 convolution blocks and 3 fully connected layers are used. The convolution block consists of a layer of Conv3x3, MaxPooling2x2, BN layer and ReLU layer, and the second convolution block does not contain MaxPooling2x2.The convolution layer acts as a feature extractor. The pooling layer can extract features, reduce training parameters and thus prevent overfitting. The main role of BN layer is to alleviate the gradient disappearance and explosion phenomenon in deep neural network training, and speed up the training speed of the model. The ReLU activation function can increase the nonlinear expression of the network. The fully connected layers play the role of classifier in the whole convolutional neural network. While the convolutional, pooling, and activation function layers map the original data to the hidden feature space, the fully connected layers weight and sum the previous features to map the learned distributed feature representation to the sample label space. The operation of the convolution block $\mathrm{CB}(\mathrm{X_R})$ is computed as

$$CB(X_R) = B(\delta(M(Conv_3(X_R)))), \tag{1}$$

where $\mathrm{X_R}$ denotes RGB images, $\mathrm{Conv_3}$ denotes 3x3 convolution operation, M indicates MaxPooling operation, B indicates BatchNorm layer and δ denotes ReLU activation function. The output feature maps of the convolution block are fed into the LSTM block which consists of a LSTM layer [21] and a fully connected layer. The operation of the LSTM block $\mathrm{LL_R}(\mathrm{X_R})$ can be expressed as

$$\mathrm{LLR(XR)=Linear(LSTM(CB(XR)))}, \tag{2}$$

The connection between the fully connection layers (FC3 and FC4) and LSTM is illustrated in Figure 2. The features from FC3 are fed into LSTM for dynamic feature extractions in time domain, and the outputs are then fed into FC4.After FC4, the features obtained from the two branches are summed to fuse the RGB image information with the depth information, and finally a fully connected layer (FC5) is used to obtain the final classification result. The 3D-DGR network output $\mathrm{Output}(\mathrm{X_R}, \mathrm{X_D})$ is computed as

$$Output(X_R, X_D) = linear(LL_R(X_R) + LL_D(X_D), \tag{3}$$

4. Results and Discussion

4.1. Data set and evaluation criterion

The experimental data set is derived from a self-built 3D dynamic gesture data set, with a total of 800 dynamic gesture color image sequences and corresponding depth image sequences in binary files, each with a size of 64x64. The data set contains 8 dynamic gestures (see Figure 5), with the RGB image of the gesture at the top and the corresponding depth image at the bottom. In the order from top to bottom, the dynamic gestures are for web navigations: minimize, backward, maximize, move up, zoom in, zoom out, move forward, and move down. In this data set, the images contain only two classes: the gesture class and the background class. The gesture class, which is the target gesture to be detected, is also called the positive sample; the background class, which is the other remaining parts, is also called the negative sample.

Figure 5: Dynamic gestures (8 kinds of actions for web navigation)

The evaluation metric uses Accuracy, a standard metric for image recognition. Theoretically, the larger the Accuracy value (the closer to 1), the better the model effect. Its calculation is expressed as

$$\text{Acc} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}, \tag{4}$$

where TP means that the predicted result is the gesture pixel and true is the gesture pixel, i.e., the prediction is correct. FP denotes prediction result is the gesture pixel and true is the background pixel, i.e., the prediction is wrong. FN means that the prediction result is a background pixel and the true is a gesture pixel, i.e., the prediction is wrong. TN denotes the prediction result is a background pixel and the real pixel is a background pixel, i.e., the prediction is correct.

4.2. Experimental environment and training details

The hardware platform for this experiment is a PC with Ubuntu 18.04 operating system, CPU is Intel Corei5-9500CPU @3.00GHZ x 6, RAM 7.6G, disk memory 100G, and Intel RealSense D415 depth camera. The experiments were conducted indoors, and the distance between the experimenter's hand and the camera was between 0.3m and 0.5m. The modeling approach proposed in this paper uses the PyTorch framework for the experiments, with 70% of the input images as the training set and 30% as the validation set. The learning rate was 0.0005, the number of batches (batch size) was set to 16, the

adaptive matrix estimation algorithm (Adam) optimizer was used [22], and the learning strategy used StepLR.

5. Results

This model experiment includes two different models for testing, the first model is learned using only RGB images, i.e., the upper branch part of 3D-DGR, and the second model, i.e., the 3D-DGR model proposed in this paper. The evaluation results of the two training processes are shown in Figure 6. It can be seen that the 3D-DGR model is superior to the model using only RGB images. When using only RGB image data, the camera cannot accurately perceive the distance of the finger from the camera, because it lacks some information to recognize the gesture more accurately.

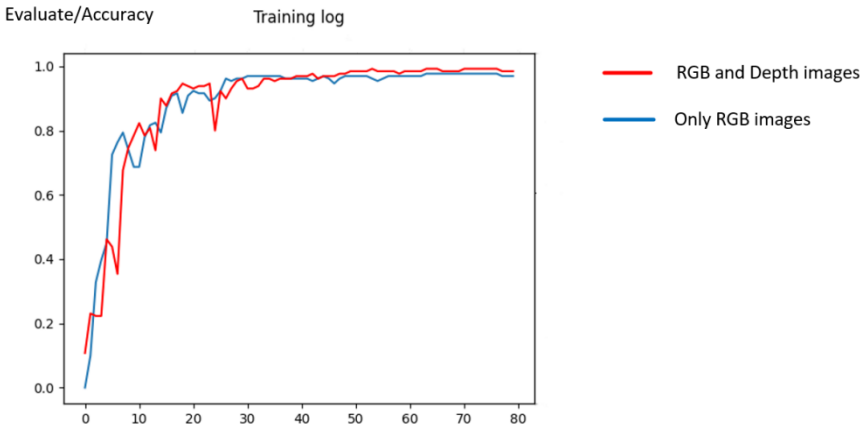


Figure 6: Accuracy of different model trainings

The exact evaluation metrics of the different algorithms is given in Table 1, and it can be clearly seen that the 3D-DGR model has a 1.52% higher accuracy than the model using only RGB images.

Table 1
Evaluation metrics of different model

Model	Accuracy	Parameters
Only RGB images	0.9771	1.3M
RGB and Depth images	0.9923	2.6M

6. Conclusions

In this paper, in order to better resolve the interference of environmental factors in the process of dynamic gesture recognition, a 3D dynamic gesture recognition network model is proposed that can fuse RGB image and depth image information using CNN and LSTM. In such a model, the RGB images and depth images captured by Intel RealSense camera are fed into CNN and LSTM to extract spatial and temporal features and then fused for gestion classification. The effectiveness of our proposed model is verified by experiments on Intel RealSense D415 depth camera. With self-built gesture dataset for web-page navigations, the achieved accuracy is 99.23%, which is 1.52% higher than that using only RGB images. Our proposed model can be used for real-time web page navigation and other useful applications.

7. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61961014 and the Hainan Provincial Natural Science Foundation of China under Grant 620RC556.

8. References

[1] Xie Yinggang, Wang Quan Overview of vision based dynamic gesture recognition [J] Computer engineering and application, 2021,57 (22): 68-77.

[2] Zhu Baozeng Gesture recognition based on depth data and its application [D] Harbin Institute of technology, 2017.

[3] Sun Bowen, Yu Feng Dynamic gesture recognition and interaction of monocular camera based on deep learning [J] Journal of Harbin University of technology, 2021,26 (01): 30-38 DOI:10.15938/j. jhust. 2021.01.005.

[4] Adithya V, Rajesh R. A deep convolutional neural network approach for static hand gesture recognition[J]. Procedia Computer Science, 2020, 171: 2353-2361.

[5] Li Y, Wang X, Liu W, et al. Deep attention network for joint hand gesture localization and recognition using static RGB-D images[J]. Information Sciences, 2018, 441: 66-78.

[6] Pisharady P K, Saerbeck M. Recent methods and databases in vision-based hand gesture recognition: A review[J]. Computer Vision and Image Understanding, 2015, 141: 152-165.

[7] Ameur S, Khalifa A B, Bouhlel M S. A novel hybrid bidirectional unidirectional LSTM network for dynamic hand gesture recognition with leap motion[J]. Entertainment Computing, 2020, 35: 100373.

[8] Raheja J L, Minhas M, Prashanth D, et al. Robust gesture recognition using Kinect: A comparison between DTW and HMM[J]. Optik, 2015, 126(11-12): 1098-1104.

[9] Carcangiu A, Spano L D, Fumera G, et al. DEICTIC: A compositional and declarative gesture description based on hidden markov models[J]. International Journal of Human-Computer Studies, 2019, 122: 113-132.

[10] Krishnan R, Sarkar S. Conditional distance based matching for one-shot gesture recognition[J]. Pattern Recognition, 2015, 48(4): 1302-1314.

[11] Joshi A, Monnier C, Betke M, et al. Comparing random forest approaches to segmenting and classifying gestures[J]. Image and Vision Computing, 2017, 58: 86-95.

[12] Nunez J C, Cabido R, Pantrigo J J, et al. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition[J]. Pattern Recognition, 2018, 76: 80-94.

[13] Ibañez R, Soria Á, Teyseyre A, et al. Approximate string matching: A lightweight approach to recognize gestures with Kinect[J]. Pattern Recognition, 2017, 62: 73-86.

[14] Hernández-Vela A, Bautista M A, Perez-Sala X, et al. Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in rgb-d[J]. Pattern Recognition Letters, 2014, 50: 112-121.

[15] Li C, Xie C, Zhang B, et al. Deep Fisher discriminant learning for mobile hand gesture recognition[J]. Pattern Recognition, 2018, 77: 276-288.

[16] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28: 91-99.

[17] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[18] Caputo A, Giachetti A, Giannini F, et al. SFINGE 3D: A novel benchmark for online detection and recognition of heterogeneous hand gestures from 3D fingers' trajectories[J]. Computers & Graphics, 2020, 91: 232-242.

[19] Dos Santos C C, Samatelo J L A, Vassallo R F. Dynamic gesture recognition by using CNNs and star RGB: A temporal information condensation[J]. Neurocomputing, 2020, 400: 238-254.

[20] Deng X, Yang S, Zhang Y, et al. Hand3d: Hand pose estimation using 3d neural network[J]. arXiv preprint arXiv:1704.02224, 2017.

[21] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[22] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.