Visual gesture recognition based on hand key points

Boxu Chena, Lixin Yu, Xiao Meng and Yang Hua
Beijing Microelectronic Technology Instiyute, Beijing, 100076, China
\textsuperscript{a}13263600636@163.com

Abstract. In recent years, computers have become a part of people's daily life, and the interaction between people and computers has increasingly become a hot spot in the field of scientific research. Gesture recognition based on vision is an indispensable part of the new generation of human-computer interaction. This paper presents a visual gesture recognition method based on hand key points, which realizes the detection of hand key points in the current input image and the recognition of defined gestures. It provides a new technical scheme for human-computer interaction application. Compared with the current mainstream solutions that directly train the defined gestures and obtain the gestures in the input image through template comparison detection, this paper first uses the rigid characteristics of the palm to design a palm detection model instead of directly detecting the entire hand. After detecting the presence of hand in the picture, the hand key point model locates 21 3D hand key point coordinates in the previously detected hand area through direct coordinate prediction. Finally, the meaning of gestures in the input image is obtained through the positional relationship between the key nodes. This solution achieves 95.7% of the accuracy of detection, and improves the FPS from 21-23 to 30-31.

1. Introduction
In the field of human-computer interaction, gestures have many application scenarios, such as game control, smart home, real-time sign language translation, etc\cite{1}. Compared with the gesture recognition based on external devices, the gesture recognition based on vision is more convenient and flexible, and reduces the equipment cost. As a natural and intuitive way of human-computer interaction, the development of vision based gesture recognition is of great significance. As early as 1993, Thomas gave the hand gestures for manual interaction using gloves\cite{2}. Although the recognition is accurate, the existence of glove peripherals is not friendly enough for users. With the development of neural networks and deep learning, the development of gesture recognition is also improving day by day, and the visual gesture recognition based on template matching is widely used\cite{3,4,5}. Researchers have also continuously optimized the accuracy and real-time of recognition. Lu di and Ma Wen qiang proposed a gesture recognition based on improved YOLOv4-tiny algorithm\cite{6}, which improved the number of transmission frames per second. Lu zhi and others designed a one-shot learning classification and recognition of gesture expression from the egocentric viewpoint in intelligent human-computer interaction\cite{7}, and realized the gesture classification under the first perspective.

However, in practical application, for each defined gesture, a large amount of data needs to be collected for model training. When new gesture information needs to be added, the model needs to be retrained\cite{8,9,10,11,12}. As a result, the requirement for input image is limited to the hand image, and the whole system becomes extremely complex due to the need to store the information of related template model.

In order to enrich the use of the scene, improve the flexibility of the system and the accuracy of detection, this paper proposes a visual gesture recognition method based on hand key points. Referring to the whole body key point detection method in Openpose\cite{13}, we similarly designed a hand key point
detection module. After detecting the key points, return their three-dimensional coordinates, and directly use the position relationship between the key point coordinates to identify the meaning of gestures in the input image. Figure 1 shows the structure of the whole system. Firstly, a palm detection model is designed based on the rigid feature of the palm to detect whether there is a hand in the current image. After the palm is detected in the current image, the hand key point model locates 21 3D hand key point coordinates in the previously detected hand area through direct coordinate prediction. Finally, the meaning of gestures in the input image is obtained by comparing the positional relationship between key nodes.

![Figure 1.](image.png)

**Figure 1.** the structure of the whole system.

2. Materials and Methods

Different from the model training for each defined gesture in the template matching method, we mainly trained the following two models: the palm detection model to detect whether there is a hand in the current image, and the hand key point detection model is used to locate 21 hand key points. After each key point is obtained, the normalization algorithm is used to calculate the horizontal and vertical coordinates of each hand node. At the same time, a spatial coordinate system is established to obtain the depth coordinates of each hand node relative to the origin of the coordinate, and finally gesture recognition is performed according to the three-dimensional coordinate position relationship of these nodes.

2.1. Palm detection model

Before detecting the key points of the hand, it is necessary to detect whether there is a hand in the image. Different from the detection targets with obvious features such as the human body and the face, the fingers are flexible and movable, which makes the feature extraction very difficult. To solve this problem, we use the rigid features of the palm to design a palm detection module as the front module of hand key point detection and calibration. The rigidity of the palm refers to the fact that the palm is an unchangeable shape in most cases, compared with the fingers that can bend and straighten.

In order to detect the position of the palm, we trained an SSD model to perform real-time detection of the current input image. The detection results are described by a square detection box. We ignored the aspect ratio of the return rectangle, which makes the network structure lighter. Lastly, we minimized the focal loss during training to support a large amount of anchors resulting from the high scale variance.

2.2. Hand key point detection model

After the hand is detected in the current input image, we use the hand key point detection model to locate the key points in the detected hand area. Figure 2 shows the selection of 21 key points of the hand. In order to facilitate the calculation of the depth coordinates, we selected the root of the palm as the 0 key
point, and the other key points correspond to the bone nodes of the hand one by one. This scheme clearly shows the activity state of each finger joint, which lays the foundation for the gesture recognition part.

Figure 2. 21 3D hand key points.

Considering that the left and right hands can collectively represent a gesture, we also added a binary classification network in the model to distinguish whether the hand in the current image is a left hand or a right hand. Therefore, the hand key point detection model has two outputs (see Figure 3):

- A binary classification of handedness, e.g. left or right hand.
- The three-dimensional coordinates x, y, z of the 21 key hand nodes.

Figure 3. The inputs and outputs of hand key point detection model.

The coordinates of the hand nodes are represented by x, y, z, and are stored in a 21*3 array. Where x and y are the normalized results on [0,1] according to the length and width of the input image, and the formula is as follows:

\[
x_\text{norm} = \frac{x_i}{\text{width}}
\]

\[
y_\text{norm} = \frac{y_i}{\text{height}}
\]

z represents the depth coordinate. The space coordinate system is established with the 0 node depth at the wrist in Figure 2 as the coordinate origin. The smaller the value of Z is, the closer the node is to the camera. The normalization method is also used to determine the depth coordinates, and the formula is as follows:

\[
z_\text{norm} = \frac{z_i-z_0}{z_{\text{cam}}-z_0}
\]
2.3. Dataset
In order to solve the two problems of hand detection and hand key point detection, we created the following data sets:

- In-house collected gesture dataset: This dataset contains 1K images. The gesture in this part of the picture is more abundant and the shooting angle is more, but its limitation is that the background environment is relatively simple and its collected from only 5 people.
- In-the-wild dataset: This dataset contains 1K images. The background environment and lighting conditions of this part of the picture are more abundant, but its limitation is that the collected gestures are relatively simple.

The above two data sets are combined to form a new data set to train the palm detection model, the hand key point detection model and the binary classification network.

2.4. Gesture Recognition
After obtaining the three-dimensional coordinates of the 21 key points of the hand, we can recognize the meaning of the gesture in the current input image according to their mutual positional relationship or the relative coordinate change between two consecutive frames. For example, for the three nodes 5, 6, and 8 shown in Figure 2, when their ordinate values are arranged in the order of 8, 6, and 5, the index finger is stretched out at this time. If the arrangement sequence is 6, 5, 8, then the index finger is bent.

3. Results & Discussion
For the palm detection model, we ignored its aspect ratio when returning to the detection box, and used a square detection box. This approach reduces the number of anchors by about three times, and improves FPS, see Table 1 for details. For the hand key point detection model, we combined the two data sets and used it. This approach avoids their limitations to a certain extent and improves the detection accuracy. The results are shown in Table 2.

| Dataset                        | Average Precision |
|--------------------------------|-------------------|
| Only In-house collected gesture dataset | 86.2%             |
| Only In-the-wild dataset       | 92.1%             |
| Combined                       | 95.7%             |

We defined ten digital gestures to verify the accuracy of recognition. Each gesture was recognized 10 times from different angles. Finally, the average recognition rate was calculated. The results are shown in Table 3. Compared with the gesture recognition method of template matching, the flexibility and scalability of ours have been improved. Specifically, when we need to add a new gesture definition, the template matching gesture recognition method needs to re-collect and train the gesture data set to be defined, and finally integrate the new model obtained into the original model. Our method only needs to add a new definition to the relative position of the node to define a new gesture information. However, compared with the template matching method to extract and detect the features of each gesture, we define gesture information through the position relationship between nodes, which requires the accuracy of hand key point positioning. In fact, considering the real-time requirements and the limitations of the data set, the average recognition accuracy of this method reaches 95.7%, which is slightly lower than 97.4% of template matching, but it is still acceptable.
Table 3. Results of two different schemes for ten digital gestures.

| Gesture meaning | Average Precision(%) | Average fps | Time to add(min) | Average Precision(%) | Average fps | Time to add(min) |
|-----------------|----------------------|-------------|------------------|----------------------|-------------|------------------|
| Template matching |                      |             |                  |                      |             |                  |
| one             | 97.6                 | 21.8        | 18               | 91.3                 | 30.1        | 1                |
| two             | 97.7                 | 21.5        | 25               | 93.9                 | 28.5        | 1                |
| three           | 94.3                 | 22.1        | 27               | 95.1                 | 30.2        | 1                |
| four            | 98.2                 | 21.6        | 22               | 98.3                 | 29.6        | 1                |
| five            | 98.7                 | 22.6        | 19               | 97.2                 | 30.7        | 1                |
| six             | 95.5                 | 22.3        | 21               | 96.3                 | 29.1        | 1                |
| seven           | 93.5                 | 20.1        | 31               | 87.7                 | 29.7        | 5                |
| eight           | 96.4                 | 22.7        | 26               | 96.8                 | 28.4        | 1                |
| nine            | 92.6                 | 21.8        | 35               | 85.4                 | 27.3        | 5                |
| ten             | 99.4                 | 23.0        | 16               | 98.5                 | 30.4        | 1                |

4. Conclusions

In this paper, we proposed a visual gesture recognition method based on hand key points. Different from the template matching method that needs to establish a data set and train related models for each gesture, we use a scheme that first locates the key points of the hand in the graph and then performs gesture recognition based on the relative position relationship between the nodes. The input image is no longer limited to a single hand, which enriches the scenes used and improves the scalability of the system.

In addition, only visual sensors are required, and no special motion or depth sensors are required. It has good reconfigurability and portability. At the same time, the relative three-dimensional coordinates of each hand node are obtained during hand key point detection. In future work, we can define the dynamic gesture through the coordinate changes of hand nodes between continuous image frames, and then realize the interactive control through dynamic gesture, such as turning pages by waving during reading. It provides a new solution for human-computer interaction.

References

[1] XIA Zhaoyang, ZHOU Chenglong, JIE Junyu, et al. (2020) Micromotion gesture recognition based on multi-channel frequency modulated continuous wave millimeter wave radar[J]. Journal of Electronics & Information Technology. 42(1): 164–172.
[2] XU Yan-xu, ZHANG Qi, WU Xia. (2013) Research of algorithm for real-time gesture recognition based on OpenCV[J]. Information Technology. (4):99-102.
[3] ZHANG Xun, CHEN Liang, HU Cheng, et al. (2017) A real-time recognition method of static gesture based on depth learning[J]. Modern Computer, (34): 6–11.
[4] PENG Yuqing, ZHAO Xiaosong, TAO Huifang, et al. (2019) Hand gesture recognition against complex background based on deep learning[J]. Robot, 41(4): 534–542.
[5] WANG Fenhuai, HUANG Chao, ZHAO Bo, et al. (2020) Gesture recognition based on YOLO algorithm[J]. Transactions of Beijing Institute of Technology, 40(8): 873–879.
[6] LU Di, MA Wenqiang. (2021) Gesture Recognition Based on Improved YOLOv4-tiny Algorithm[J]. Journal of Electronics & Information Technology:1-9.
[7] LU Zhi1 QIN Shi-Yin, LI Lian-Wei ZHANG Ding-Hao. (2021) One-shot Learning Classification and Recognition of Gesture Expression From the Egocentric Viewpoint in Intelligent Human-computer Interaction[J]. ACTA AUTOMATICA SINICA, 47(6):1284-1301.
[8] Redmon J, Divvala S, Girshick R, Farhadi A. (2016) You only look once: Unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE. 779–788
[9] Zhang Y F, Cao C Q, Cheng J, Lu H Q. (2018) EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition. IEEE Transactions on Multimedia, 20(5): 1038–1050

[10] Long Zhang, Lu Chengya, Li Guopeng, Zhang Weilie, Wen Feijuan, Li Bo1 (2021). Design of Mobile Robot Gesture Control System with Visual Gesture Recognition. Mechanical Science and Technology for Aerospace Engineering, doi:10.13433/j.cnki.1003-8728.20200449.

[11] Tian Q H, Yang H M, Liang Q L, et al. Overview on vision-based dynamic gesture recognition[J]. (2020) Journal of Zhejiang Sci-Tech University, 43: 1-13

[12] Zhang B Y, Zhong Y, Li Z D. A visual object tracking algorithm based on dynamics pattern and convolutional feature[J]. (2019) Journal of Northwestern Polytechnical University, 37(6): 1310-1319

[13] Qiao S., Wang Y., & Jian L. (2018) Real-time human gesture grading based on OpenPose. International Congress on Image & Signal Processing. IEEE.