Research on hand key point recognition and localization method based on SSD network

Xiaoyu Cheng1,a, Tao Jiang2,b, Xuewen Tan3,c, Lidong Huang4,d

1School of Mathematics and Computer Science Yunnan Minzu University Kunming, China
2School of Mathematics and Computer Science Yunnan Minzu University Kunming, China
3School of Mathematics and Computer Science Yunnan Minzu University Kunming, China
4School of Mathematics and Computer Science Yunnan Minzu University Kunming, China

a18987461960@163.com, bJtzwyinternet73@163.com, c657465359@qq.com

d* Corresponding author: hldinter@163.com

Abstract. With the development of the times and the progress of society, multimedia presentation has become a necessary skill for people. In the field of education, the application of multimedia presentations for teaching activities has become a necessary and common way. In this paper, we propose to use camera capture for image analysis and wireless data transmission technology for information interaction with computers. Among them, SwiftShow performs image analysis by capturing the gestures of the presenter using the camera. The device provides personalization, and users can choose the mode that suits them through the interaction interface, and even customize the gestures. The device can be controlled by the simple hand gesture recognition overhead, which not only can realize the basic functions such as ppt play, pause, page turn, and focus, but also includes many practical and unique presentation functions, the research aims to completely free the presenter from the computer.

1. Introduction

In the rapid application of multimedia presentation, the initial gesture recognition is mainly using mechanical equipment that directly detects the angle and spatial position of each joint of the hand and arm. These devices mostly interconnected the computer system and the user through wired technology, so that the user's gesture information was transmitted to the recognition system completely and without error, with typical devices such as data gloves. Subsequently, the optical marker method replaces the data glove to wear the optical marker on the human hand, and the change of human hand position and finger can be transmitted to the system screen through infrared. This method can also provide good results, but still requires more complex equipment [1]. Although the intervention of external equipment makes the accuracy and stability of gesture recognition improve, it obscures the natural expression of the gesture For this reason, the visual-based gesture recognition method was born. Visual gesture recognition refers to the processing of image sequences containing gestures captured by video capture devices through computer vision technology to recognize the gestures.
2. Working Principle

Hand gesture recognition is the process of classifying trajectories or key points in the model parameter space to a subset in that space, and we propose a method for hand key point recognition and localization based on SSD networks. It is difficult to obtain accurately labeled data sets for human hands in 3D space due to different viewing angles and flexible fine movements. We first train the neural network with a small number of labeled datasets containing human hand keypoints, initially detect the keypoints with the detection model, get the 3D positions of the keypoints, then reproject the calculated 3D point positions to each 2D image with different viewpoints, and then train the detection model network with these 2D images and keypoint labeling, after several iterations, that is, we can get a more accurate hand keypoint detection model. The model can generate 22 key points, 21 of which are for the human hand and the 22nd point represents the background. The following figure shows the locations of the 21 keypoints on the human hand [2].

![Diagram of the key points of the hand.](image)

The image of the hand is passed into the DCNN network for feature extraction to locate the joints of the human palm. This method can better portray the pose of the hand than the neural network that recognizes the hand contour. Determining the joint position of the lower palm and the afferent analysis module can analyze the corresponding hand posture. Our neural network model uses a model migration approach to retrain the output layer of the SSD network using google's hand localization model. Since the original neural network model is trained from a huge hand data set, it has a good recognition capability. We migrated a joint localization function on top of it and experimentally proved its accuracy.

In order to make the model achieve the recognition accuracy shown now, we conducted a lot of experiments during the design process and used the SSD neural network structure to solve the key point jitter problem. The network layers after the VGG model layer used in the first two layers of this structure network use four convolutional layers and four pooling layers. The incoming images to our network are image frames captured in real time and used to extract shallow features. Then, we use three Inception modules and two Reduction modules to deepen the network [3]. Finally, we combine the center loss with the softmax loss to make the network more representative and finally locate the planar coordinates of the hand joints. For the problem that the key points have jitter and some frames may have key point jumps, we use Savgol filter to smooth the data in order to make the recognition labeling results more stable, and we can get more satisfactory results.

3. Solution Concept

3.1. Functional design objectives of the system

- Connection to the host: the receiver will be inserted into the USB port of the computer host, the signal receiver will automatically install the driver, using 2.4G wireless data transmission technology to connect and communi-cate for the control signal reception of the presentation.
- Hand motion capture: The device captures the video information of the user's hand motion in real time through a high-definition camera. The captured image information is supplied to the hardware and drivers for processing.
- Presentation display control: control equipment through the gesture recognition algorithm analysis of the captured gesture image, according to the real-time recognition of the hand command to complete the corresponding display function, complete such as start screening,
click trigger, ppt page flip; in the presentation process may be carried out in the music, video playback software simple control such as start, stop, pause, program page switch.

- User-defined mode and gestures: the device can switch the usual mode through the UI interface, while customizing the mode and combining the gesture library and command library by itself; resetting the corresponding gestures of the command, the user can enter the gestures, the system will analyze the recognition difficulty of the new gesture to judge the feasibility, if it does not pass, the user will be asked to re-enter, and if it passes, the data will be updated.
- Device management: The device has function prompt lights to show the current connection of the device. Users can use the device's buttons to achieve common functions such as mode switching, and the function of the buttons can be changed in the UI interface.

### 3.2. Overall design

Each user of the multimedia presentation device can be assigned a customized gesture presentation mode. The core of the gesture recognition control presentation process is to extract key point location information from the user's specific gestures, and after wireless communication, to functionalize the gestures through API programming. When the user selects the custom mode, the user is free to pair the function library with the gesture library [4]. When the alternative options in the gesture library cannot meet the user's needs, the user can select a custom gesture to call the core function module for gesture entry, and the algorithm compares the gesture features stored in the device to make the entry permission judgment. The overall design diagram is shown in the following figure.

![Overall design diagram of the program.](image)

The main body of the multimedia presentation system design based on hand action recognition technology lies in the recognition and judgment algorithm of hand keypoints and the high-speed response device. A dynamic key point recognition method based on DCNN (Deep Convolutional Neural Network) is proposed in the recognition algorithm, and the device contains both hardware and embedded software. The hardware side needs to write and embed drivers for the wireless data receiver of the Unicom host, and call Opencv to intercept the input signal of the mouse for simulation to realize the control of the presentation and other programs. Hardware devices such as cameras, also need to respond to the driver, the software side of the two DCNN models, consisting of five convolutional layer groups and three fully connected layers, where each convolutional layer group consists of a convolutional layer and a pooling layer for training and testing respectively. The following is a schematic of the software part of the design as shown below [5].
3.3. **Detailed design**

3.3.1. **Ppt Display:** The core function of the device is ppt display and custom gestures. Ppt display process design is: user turns on the power and plug the signal receiver into the computer, the device will automatically detect the connection to control the host, and start the video capture program, the relevant driver will capture the video data to the image pre-processing program and capture the image frame at the specified interval, and then passed to the model for key point identification markers are then passed to the model for key point identification. The gesture analysis module translates the action of the key point inscription in a time sequence to get a fixed command and sends the command to the receiver to execute the relevant operation.

![Figure 3](image_url)  
Figure 3 Design schematic of the software part.

3.3.2. **Gesture Customization:** The general process of gesture customization is: when the signal receiver is connected to the host for the first time, the driver will be loaded automatically. After the connection is established, personalized settings can be made through the UI interface provided by the successfully installed plug-in. If no changes are needed, the gesture can be demonstrated directly in the default mode or the current mode. If changes are needed, the user selects a mode, clicks to select it, and double clicks to enter the display interface of that mode to view the gesture information contained therein, as well as to view a video of the gesture demonstration instructions. If you want to change it, you can select a

![Figure 4](image_url)  
Figure 4 User PPT presentation flow chart.
mode, click on it, and double-click on it to go to the display of that mode, view the gesture information contained in it, and view the demonstration video of the gesture. If the user needs, he/she can choose to add a new mode, each mode is assigned no more than 12 instructions, and the user can select the gesture library one by one to match. If the selected gesture is more repetitive than the previous gesture (more likely to cause recognition errors) the system will undo the self-defined gesture and give the user a hint, if the evaluation is within the feasible range, the system will enter the gesture to indicate the corresponding command. If the recorded gestures in the gesture library do not meet the user's needs, the user can create new gestures in the gesture entry module to update the gesture library [6]. The device will open the video capture program, pass the captured video data to the image pre-processor through the relevant driver and capture the frames of the image at the specified interval, and then pass them to the model for key point identification marking. The combination of coordinates of the new keypoints and the existing valid gesture keypoints will be evaluated for feasibility, and if the evaluation passes, the update is successful and the user can use the new gesture to communicate commands.

4. Results Show

4.1. Hardware Implementation
In this work, we designed and developed a new type of remote control multimedia office device using gesture recognition in order to investigate the ability to capture the gestures of the displayer using a camera for image analysis. Figure 5 shows the 3D model, and Figure 6 shows the internal mechanical structure of the display.

![Figure 5 Multimedia office equipment model diagram.](image1)

![Figure 6 Multimedia office equipment internal structure diagram.](image2)

Part of the device is a USB signal receiver connected to the host computer. The USB interface has the characteristics of hot-plugging and plug-and-play, and is very easy to use, so it makes sense to design
a wireless data reception system based on the USB interface. Traditional cable communication has the
difficulty of wiring, and the USB interface technology has the advantage of supporting hot-plugging.
Therefore, we designed a USB-based wireless link system, which can transmit the data received by the
wireless receiver module to the PC through the USB interface, complete the effective data transmission
between the PC and the data generator, and can complete the operation function of the data generator to
the PC's multimedia presenta-tion. This design uses ADI's high performance microcontroller AduC845
as the main controller to realize the data received by the wireless transceiver chip nRF2401 through the
USB chip CH375 into a data packet with USB protocol and transmit it to the PC to operate the PC.

The ARMCortexA8, or A8 or CortexA8 for short, is the first application processor based on the
ARMv7 architecture, using Thumb-2 technology that delivers higher performance, power efficiency and
code density. The Cortex-A8 the first application processor based on the ARMv7 architecture with a
main frequency of 600MHz to 1GHz, NEON signal processing extension set, providing acceleration
for media codecs such as H264 and MP3, and power consumption below 300 milliwatts.

CCD camera surface is covered with silicon semiconductor photosensitive elements to capture
photons and generate photogenerated electrons, these electrons are first accumulated in the insulating
layer under the CCD, and then exported by the control circuit in a serial manner to the analog-to-digital
circuit, and then through the DSP and other imaging circuits to form the image. The biggest difference
between fast scan and slow scan is that the speed of photogenerated electrons exported and the circuit
system is different. Fast scan exported electrons at a very fast frequency in order to achieve a video-level
refresh rate, but this will lead to the loss of electrons, increased noise, photogenerated electron emptying
is not complete; while slow scan is the opposite, its circuit design focuses on the protection of photogenerated
electrons On the contrary, slow scan, its circuit design focuses on the protection of photogenerated
electron accumulation, the export frequency is not high, but to ensure that the process of electronic loss
and loss to a minimum, its analog-to-digital converter dynamic range and sensitivity is very high, to
ensure that the signal conversion process is not distorted, while in order to reduce the thermal effect of
noise, the general use of Cooling system to cool down.

4.2. Software Implementation

After the preliminary targeted research, based on the clarification of the working principle, through the
overall design, detailed design, hardware selection and computer programming, we finally realized
many functions such as home page mode selection, mode gesture list display, mode gesture display,
custom mode and custom gesture, etc [7]. The result satisfies the initial prediction. The specific results
are shown in the following figure.

![Figure 7 Ideal gesture recognition results.](image-url)
5. Conclusion
With the rapid development of deep learning technology, multimedia presentation technology has made great progress and become a research hotspot in computer application field in recent years. In this paper, the research objectives, research background and current situation, working principle and program vision are all elaborated, and mainly by using the camera to capture the gestures of the presenter for image analysis, the user can choose the suitable mode through the interactive interface, and even customize the gestures. This research can effectively meet the practical needs of the presenters, and it is an original implementation of the function of writing and marking with gestures, which can solve almost all the problems that have arisen in the presentation process and eliminate the rigid and unbreakable limitations of the traditional laser pointer. In addition, the research can also adapt to the current needs of multimedia speech presentation, maximize the use of personnel speech effect, improve the use of modern information tools and work efficiency, aimed at teaching staff and business people to provide the most ideal intelligent presentation program.

References
[1] Liao Y, Yang J, Wang Z, Wang S. Adaptive migration learning based on bimodal convolutional neural network for flotation condition identification[J]. Journal of Photonics, 2020, 49(10): 173-184.
[2] Yang Y, Fan D, Yin H, Han J, Miao S. Deep-migration learning based transmission line fault phase selection model and its migrability[J]. Power Automation Equipment, 2020, 40(10): 165-172.
[3] Dong S, Zhao S. Research on human-computer interaction technology based on gesture recognition[J]. Electronic Production, 2020(19): 73-74+54.
[4] Li L, Xi S, Zhang C, Ma H. An algorithm for workpiece surface defect identification based on improved SSD model[J]. Computer Engineering and Science, 2020, 42(09): 1608-1615.
[5] Fu T, Yu L. Improved convolutional neural networks for dynamic gesture recognition[J]. Computer Systems Applications, 2020, 29(09): 225-230.
[6] Wang Q, Fan D, Li S, Zhang Shisen. Time-series action localization based on dual-stream convolutional neural network[J]. Software Guide, 2020, 19(09): 35-38.
[7] Jiang X. Research on hand pose estimation method based on point cloud attention model[D]. Dalian University of Technology, 2019.