Real World Hand Gesture Interaction in Virtual Reality

To cite this article: Qi Zhang et al 2019 J. Phys.: Conf. Ser. 1229 012027

View the article online for updates and enhancements.
Real World Hand Gesture Interaction in Virtual Reality

Qi Zhang*, Wenzhe Zhu and Qing Zhu
Faculty of Information Technology Beijing University of Technology, Beijing, China

*lovefuji.ok@163.com

Abstract. With the development of virtual reality, the appropriate interaction technology has become the focus of practitioners. Today's virtual reality interaction is mostly through the interaction of traditional electronic devices, such as handles and other products, temporarily solve the problem of interaction, but also make the user from the virtual world is taken out, immersion greatly reduced. The paper uses hand gestures to interact with virtual reality. By using skin detection in YCbCr color space and indicates the location of hand to determine the approximate range of hand. Then, the histograms of oriented gradient (HOG) which has been widely used in target detection research in recent years is used to extract gesture features, and applies support vector machine (SVM) method to achieve the real-time hand gesture recognition. Finally, experiments prove that the proposed method is accurate and stable in virtual reality interaction.

1. Introduction
Unlike other technology products, virtual reality [1] (VR) emphasizes immersion, and the immersion is created by isolation from the outside world, especially visual and auditory. Therefore, VR can deceive the brain and try to get us to hold the notion that we are in the real world. However, in virtual environment, people can’t see the body and interact with the virtual environment by body-gestures. Traditional electronic devices, such as mouse, keyboard[2], handles, temporarily solves the problems of interaction in virtual environment, but it also brings users out of the virtual world, which greatly reduces their immersion.

In this work, we focus on the improvement of the interaction method in virtual environment. Our goal is to recognize hand gestures to interact with virtual reality. The experiments are carried out under challenging, characterized by inoculating hand which segment from complex background and virtual environment, and running in a real-time environment.

2. Image Segmentation
In this work, we segment hand image from complex background. In practical applications, some flexible methods are used to hand gesture recognition, such as restricting the background color and light intensity, or covering the user's hand with a specific color. This is very unfriendly in the sense of use, so a method which is insensitive to light and background color is needed to separate the hand image to achieve the purpose of putting the real hand influence in the virtual scene.

First, the image is converted to YCbCr color space for skin color segmentation [3] to determine the approximate contours of hand. Then we put this range into a Mask image. The mask image is removed small color patches which not belongs to hands by morphology. After that, use the mask image to segment hand image from complex background.
2.1. Skin detection
Skin color model is a mathematical model describing the distribution of skin color. At present, there are several commonly used static skin color models: threshold method, simple Gaussian model, mixture Gaussian model, histogram statistics and region-level detection. It can be clearly seen that the threshold method is better than other methods in response speed, which is more conducive to the use of real-time systems. Experiments show that in YCbCr color space, the clustering characteristics of human skin color are better, which is conducive to skin detection [4].

The image from video which token by monocular camera is converted to YCbCr color space for skin color segmentation.

In the YCbCr color space, the Y component represents the brightness. Cb and Cr components represent the blue components and the red components, that is, colors. In another word, the characteristic of YCbCr space is to separate chromaticity and brightness with Y and CbCr. Experiments show that it’s better for cluster characteristics of human skin color in YCbCr color space. Therefore YCbCr color space is conducive to skin color detection. The transformation formula from RGB color space to YCbCr color space is as follows.

\[
\begin{bmatrix}
Y \\
Cb \\
Cr
\end{bmatrix} = \begin{bmatrix}
16 \\
128 \\
128
\end{bmatrix} + \begin{bmatrix}
0.257 & 0.504 & 0.098 \\
-0.148 & -0.291 & 0.439 \\
0.439 & -0.368 & -0.071
\end{bmatrix} \begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\] (1)

Experiments show that the color range of the hand is [0,255], [77,127], [133,173], respectively. Pixel-by-pixel inspection of the image converted to YCbCr color space is carried out. If it falls into the prescribed color range of the hand, the hand is judged as the hand, otherwise it is judged as the background.

2.2. Hand Contour Extraction
Mask which is a binary image is needed to segment the hand image from background, the mask is shown in Figure 1. (b). Value 0 (black) represent background, and value 1 (white) represent foreground, that is, the hand image. The image obtained by normal skin detection is still noisy due to environmental reasons such as illumination.

Use Median Blur to smooth the image and remove some of the noise that still exists after skin detection. Morphological methods are then used to reduce "small black holes" in the hand image. The binary operation image is used to remove the isolated noise due to illumination, and then the binary image is processed using the closed operation operator to remove the "black hole" in the hand image. Next, use the mask to cover the background portion of the original image to black. In another word, the area which value is labeled by 1 of the mask is maintained, and 0 of the area is cleared. The final hand image is presented in the virtual scene, and the origin picture and segmentation result are shown in Figure 1. (a), (c).

![Figure 1. Hand image segmentation result.](image)

3. Hand Gesture Recognition
In this section, we first look for a part of the open source first person data set and make a part of the data
set, combining the two to form the data set we want to use. Then, using HOG [5] to describe the gesture feature, the HOG feature does not need to filter out the illumination factor in the pre-processing process, and can also adapt the image deformation well, which is beneficial to the realization of engineering. Then, the gesture is learned using a support vector machine (SVM).

3.1. Dataset
This paper mainly solves the problem of seeing hands in the real world in a virtual scene and using a series of hands to interact. The gesture recognition performed in such a specific scene recognizes the gesture in the first perspective, and thus the required data set is the gesture image in the first perspective. The York University public 11k Hands dataset [6] is used here. There is only one gesture in this dataset, which is a five-finger gesture, so we need to make some datasets ourselves.

Five volunteers recorded video of the corresponding gesture for about one minute under different lighting conditions, and these gestures were gestures of the first perspective. We take the picture frame by frame and make the data set we need for our experiment.

3.2. Gesture Feature Extraction
The Histogram of Oriented Gradient (HOG) feature is used to calculate and count the gradient direction histogram of the local region of the image to form the gesture features required by our experiments.

The HOG feature does not consider features from the whole, but describes the local features of the image. A gradient direction histogram in each cell is calculated by subdividing the image into a plurality of cells. A plurality of connected cells form a block, and the contrast is normalized in the block, and finally the extracted HOG features are obtained.

The HOG features have no rotation and scale invariance, so the amount of calculation is small, adapting to the real-time environment, and at the same time being able to balance accuracy. Since the HOG feature is based on local image extraction, it can better adapt to changes in geometry and illumination. Although the rotation adaptation is slightly weaker, combined with our experimental environment and goals, the camera performs first-person gesture recognition at the human eye position, and the direction of the hand is approximately the same. Compared with the Hu invariant moments [7] commonly used for gesture recognition, the advantage of the HOG feature is that it adapts well to illumination changes and does not need to be converted into a binary map, which simplifies the processing of the data set and facilitates the optimization of subsequent projects. Therefore the HOG feature is particularly suitable for our experiments.

- Convert an input RGB image of 64x64 pixels into a grayscale image.
- Gradient calculation. The convolution operation is performed by traversing the image using the gradient gradient template [-1, 0, 1], [-1, 0, 1] T in the -x and -y directions, respectively. The gradient of each pixel in the horizontal and vertical directions is as shown by the formulas (2) and (3).

\[
\text{gradient}_x(x, y) = H(x + 1, y) - H(x - 1, y) \tag{2}
\]

\[
\text{gradient}_y(x, y) = H(x, y + 1) - H(x, y - 1) \tag{3}
\]

Then the gradient value of the pixel point \((x, y)\) and the gradient direction are expressed as shown by the formulas (4) and (5).

\[
\text{gradient}(x, y) = \sqrt{\text{gradient}_x(x, y)^2 + \text{gradient}_y(x, y)^2} \tag{4}
\]

\[
\alpha(x, y) = \tan^{-1} \left( \frac{\text{gradient}_y(x, y)}{\text{gradient}_x(x, y)} \right) \tag{5}
\]

- Construct a gradient direction histogram. The image is decomposed into several cells of 8x8 pixel size, and the gradient direction divides 360° into 9 intervals, and linear interpolation voting is performed to obtain a gradient histogram of each cell, that is, a 9-dimensional feature.
- Divide the block and normalize it in the block. In this paper, a region with a size of 2x2 cells is
connected as a block, and the step size of each sliding is 8 pixels. Avoid the difference in the gradient range in the same picture due to the influence of illumination. It needs to be normalized in block units. The L2 regularization method is used in this paper.

\[ v = \frac{v}{\sqrt{e^2 + ||v||^2}} \]  

- Obtain the HOG feature of the image. The feature vectors normalized by all blocks are concatenated as the input data required for the subsequent training recognition portion. The dimension of the feature vector of a picture in this paper is 1764.

3.3 Training and Identification

Using SVM as a classifier for training, this paper uses a Gaussian radial basis kernel function, and the kernel function is shown as formula (7). The use of the already segmented hand image displayed in the virtual scene for recognition reduces the redundant information of the image and effectively improves the accuracy of the recognition.

\[ K(x, x_i) = \exp\left(-\frac{||x-x_i||^2}{\delta^2}\right) \]  

4. Experiment and result analysis

4.1. Implementation of the experimental system

The modules involved in the experimental system are mainly divided into five parts: image acquisition module, hand image segmentation module, data training module, gesture recognition module, and business logic module. The image acquisition module is composed of a monocular camera, and the acquired image is used as input data of the next step.

- A hand image segmentation module that uses a skin color detection and morphological method to segment a hand image from a complex environment and display it into a virtual scene.
- The data training module classifies the extracted HOG features using SVM, and saves the classification result as xml as a basis for gesture recognition.
- The gesture recognition module extracts the HOG feature and identifies the hand image that has been segmented by the classification result saved by the data training module.
- Business logic module that interacts through gesture recognition in a virtual scene.

The process is shown in Figure 2.

![Figure 2](image2.png)

**Figure 2.** Experiment process.

4.2. Analysis of results

This article defines three gestures, as shown in Figure 3. Each gesture is selected as a training set and test set in a ratio of 7:3. For specific data set information, please refer to section 3.1 Dataset.
Figure 3. Hand gestures chose in this paper.

The recognition accuracy, segmentation response time, recognition response time and the average of each hand gesture are shown in the table 1. From this table, the average recognition rate of the gesture is 95.5%, the average segmentation response time is 115ms, and the average segmentation response time is 17.9ms. The result show that our system can run in the real-time environment.

Table 1. The segmentation and recognition result.

| Gesture             | 1     | 2     | 3     | Average |
|---------------------|-------|-------|-------|---------|
| Accuracy            | 89.5% | 97.8% | 99.3% | 95.5%   |
| Segmentation response time | 113ms | 116ms | 117ms | 115ms   |
| Recognition response time | 17.3ms | 17.6ms | 18.8ms | 17.9ms  |

The effect in the experimental system and the meaning of several gestures in this system is shown in Figure 4.

Touch

Hold
5. Conclusions

This paper implements a method of using gestures to interact in a virtual environment. There is no special requirement for the green screen in the real environment, and the hand image segmentation is directly realized, and the HOG feature and the SVM classification are used. This method has better adaptability to illumination and is suitable for the first person gesture recognition with a small gesture rotation angle, and conducive to the realization of engineering. Experiments show that this method is accurate and stable in virtual scenes and can run in real-time requirements.

References

[1] G. Burdea&C. Philippe, Virtual reality technology[M], Presence: Teleoperators and virtual environments, 2003, 12(6): pp.663-664.

[2] PAVLOVIC V, SHARMA R, HUANG T S. Visual interpretation of hand gestures for human-computer interaction a review [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1997, 19(7):677–695.

[3] Jochen Triesch, Christoph vnder Malsburg, A System for Person-Independent Hand Posture Recognition against Complex Backgrounds, IEEE Transaction on Pattern Analysis and Machine Intelligence , Dec. 2001,Vol . 23, No. 12, pp. 1449–1453.

[4] Shih F Y, Cheng S. Automatic seeded region growing for color image segmentation[J]. Image & Vision Computing, 2005, 23(10):877-886.

[5] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[M] CVPR, 2005.

[6] Mahmoud Afif, 11K Hands: Gender Recognition and Biometric Identification Using a Large Dataset of Hand Images, arXiv preprint arXiv: 1711.04322 (2017)

[7] Yun Liu, Zhijie Gan, Yu Sun. Static Hand Gesture Recognition and Its Application based on Support Vector Machines [C] Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Phuket, Thailand, August 2008. IEEE, 2008.