Gesture Recognition Based on the Image Acquisition Belt

Han Zhao ZHU, Yu Kun BO, Zhi Li XUE, Jia Cheng LI, Ming Hui SUN
College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China
echozhz@126.com

Abstract. Gesture recognition provides a more natural method for interacting in the field of human-computer interaction. The current popular VR glasses either use a handle as the device to interact with the virtual scene, or integrate the gesture recognition system into the VR glasses. The former mode is inconvenient and unnatural, while the latter has poor stability due to the awkward position. The input devices for the current vision-based sign language recognition systems usually have fixed positions, which makes it impossible for users to use them anytime and anywhere. Therefore, this paper proposes a gesture recognition method based on the image acquisition belt which is fastened on the user’s waist, so as to improve the user friendliness and obtain stable gesture information.

1. Introduction
Gesture recognition is a hot research area in recent years with the progress of technology. And its applications have dramatically increased. There are many ways to recognize gestures, in the early years, there are many research based on data gloves. However, today’s research has shifted to vision-based recognition due to data gloves’ high price and poor user experience. Gesture recognition based on vision has lots of better features such as non-contact, natural to use and so on.

At present the mainstream VR glasses provides either no any extra devices or only handles for interaction. To use a handle is not a user friendly interaction method for it needs users to take it with hands. At the same time, vision-based gesture recognition provides a non-contact method. As such, vision-based gesture recognition is a good alternative.

According to the different image acquisition devices, gesture recognition can be divided into RGB camera based gesture recognition[1] and depth camera based gesture recognition[2]. Gesture recognition can also be divided into static gesture recognition[3] and dynamic gesture recognition[4].

In this paper we only consider static gesture recognition and the equipments we use are RGB cameras. We also designed an image acquisition belt which can improve user friendliness significantly.

2. Proposed Methodology
In our work, vision-based gesture recognition is divided into five stages:

(1) Image acquisition, (2) Hand segmentation, (3) Feature extraction, (4) Recognition on single frame, (5) Result composition

The model of our project is shown in Figure 1.
2.1. Image Acquisition

We designed an image acquisition belt for the purpose of obtaining user’s gesture information. The equipment is composed with a belt, two wide angle cameras and braced structures. Due to the belt is fastened on the user’s waist, gesture information acquisition is relatively stable. It is possible for the user to move around without worrying about getting out of the visual field of the detection device.

Figure 3  Image before and after Calibration and Cutting
2.2. Hand Segmentation

The main purpose of hand segmentation is to detect and extract the hand region from the given image. Whether the hand region is extracted correctly and precisely determines the success or failure of gesture recognition.

![Diagram of Hand Segmentation Process]

Figure 4  Process of Hand Segmentation

The method based on skin color is widely used due to its convenience and high efficiency. Since skin color has an obvious clustering feature in YCrCb color space[5], it works fine to threshold the YCrCb image with skin color.

After thresholding the YCrCb image, we get a binary image. In this image, the pixel value of skin area is 255 while other area is 0. Then we search for contours in this binary image and the biggest contour we find is regarded as the hand region. A mask can be generated according to the contour. The mask image is also a binary image, in which the value is 255 for pixels inside the contour, 0 for others. The background is eliminated after masking the original RGB image.

We then calculate the minimum bounding rectangle of the hand contour to segment the hand region.

Since HOG can be easily influenced by image rotation, affine transformation is indispensable. Our rotation angle calculation method based on the global HOG is described below:

1. Use the entire image as a cell to compute HOG features. The number of orientation bins we choose here is 9, so a 9-dimensional vector $\alpha$ is obtained.
2. Let $\varepsilon$ be the maximum element of $\alpha$.
3. Subtract each element of $\alpha$ from $\varepsilon$. The result is represented by $\beta$.
4. Let row vector $\gamma$ be the equal of $[4, -4, -3, -2, -1, 0, 1, 2, 3]$.
5. Calculate the angle to rotate as below:

$$
\theta = \frac{\gamma \cdot \beta}{\|\beta\|} \times \frac{2\pi}{9}
$$

(1)
2.3. Feature Extraction

HOG (Histogram of Oriented Gradients) is a popular feature descriptor in computer vision field. Usually it is applied to detect objects. This method was originally raised in 1980s, but not widespread until 2005 when Navneet Dalal and Bill Triggs applied HOG in human detection[6]. Soon the method was introduced to many other types of object detection, such as traffic sign detection[7], face recognition[8], gesture recognition[1].

The process is described below[6]:
1. Normalize gamma and color of the input image.
2. Compute gradients.
3. Weighted vote into spatial and orientation cells.
4. Contrast normalize over overlapping spatial blocks.
5. Collect HOGs over detection window.

In this paper, our cell size is $8 \times 8$. The block size is $16 \times 16$ and block stride is $8 \times 8$. the number of orientation bins is 9.

2.4. Recognition on Single Frame

SVM (Support Vector Machine) is a supervised learning model that used for classification and regression analysis. Before classify data points, the SVM needs to be trained at first. The training includes two data sets, one for positive samples and the other for negative samples. Each data point is a $p$-dimensional vector which means a list of $p$ numbers. After training, a $(p-1)$ dimensional hyperplane would be found out, which can separate positive data points from negative data points with the maximum margin.

For those linearly non-separable occasions, the kernel trick is introduced. A kernel trick maps the origin data points into higher dimensional space to get a hyperplane with the maximum margin. The most frequently used kernel functions includes Linear, Polynomial, RBF (Radial Based Function) etc.
We trained a separate SVM for each gesture and camera. Then these classifiers were cascaded into 2 classifiers, each for a camera. From our experiments, there is no conspicuous accuracy gap with different kernel functions, but the linear kernel function has shortest training time and highest efficiency among all tested kernel functions. Therefore, we adopted linear kernel function.

2.5. Result Composition
Since we utilize two cameras in our system, there are two images being captured and processed at the same time. To acquire a more reliable result, we applied a statistical method to every 25 consecutive frames from both cameras. The gesture with a probability greater than 0.6 in given 25 frames is considered to be the recognition result of this camera. If there is no gesture’s probability greater than 0.6, then the result is unknown. Two gesture recognition result can be obtained through two cameras in the system. The final result would be unknown if the two results are different.

3. Experiments and Results
In our experiments, we used 400 images as positive samples and 1600 images as negative samples for each gesture during our training. The negative samples of a specific gesture includes positive samples of other gestures. We used 20 short videos as testing samples. Each video includes 25 frames. The result is shown as below (gray shadowed cells represent true positive):

| Table1 | Accuracy on Single Image of Camera 1 |
|--------|--------------------------------------|
| Acc(%) | Gesture1    | Gesture2 | Gesture3 | Gesture4 | Gesture5 | Unknown |
| Gesture1 | 73.35 | 0.55 | - | - | - | 16.10 |
| Gesture2 | 99.45 | - | - | - | - | 0.55 |
| Gesture3 | - | 75.75 | 4.89 | - | - | 19.36 |
| Gesture4 | - | 99.62 | - | - | - | 0.38 |
| Gesture5 | - | - | - | - | - | 99.65 |

| Table2 | Accuracy on Single Image of Camera 2 |
|--------|--------------------------------------|
| Acc(%) | Gesture1    | Gesture2 | Gesture3 | Gesture4 | Gesture5 | Unknown |
| Gesture1 | 100.00 | - | - | - | - | 0.37 |
| Gesture2 | 99.63 | - | - | - | - | 0.37 |
| Gesture3 | 0.19 | 98.68 | - | - | - | 0.94 |
| Gesture4 | 1.50 | - | 98.50 | - | - | 0.19 |
| Gesture5 | 5.42 | - | - | - | - | 94.23 |

| Table3 | Composed result of two cameras |
|--------|--------------------------------|
| Acc(%) | Gesture1    | Gesture2 | Gesture3 | Gesture4 | Gesture5 | Unknown |
| Gesture1 | 90.00 | - | - | - | - | 10.00 |
| Gesture2 | 100.00 | - | - | - | - | 10.00 |
| Gesture3 | 95.00 | - | - | - | - | 5.00 |
| Gesture4 | - | 100.00 | - | - | - | 10.00 |
| Gesture5 | - | - | - | - | - | 10.00 |

Gesture 1 to 5 corresponding to the following gestures:

![Gestures Used in Our Experiments](image)
4. Conclusions
In this paper, we proposed a method for 5 defined gestures recognition using image acquisition belt, HOG feature, SVM and a statistical method. We also proposed a method of rotating images to overcome HOG’s lack of rotation invariance.

Our work provides an approach to help improve the experience of interacting with VR glasses. The system in this paper can only recognize 5 gestures, further work is needed to recognize more gestures, and add other indispensable functions to achieve a practically available system.

The corresponding author is Ming Hui SUN. This work is supported by Undergraduate Training Program for Innovation and Entrepreneurship of Jilin University and Special Funds for Industrial Innovation in Jilin Province (2016C035)

References
[1] A.J, R.A, R.C, O.N and A.P. Gesture Recognition Using Support Vector Machine. International Journal of Electrical, Electronics and Data Communication, 2015, ISSN: 2320-2084.
[2] Nyirarugira, Clementine, Hyo-rim Choi, and TaeYong Kim. Hand Gesture Recognition Using Particle Swarm Movement. Mathematical Problems In Engineering 1-8. Academic Search Complete, 2016.
[3] Yeo, Hui-Shyong, Lee, Byung-Gook, Lim and Hyotaek. Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware. Multimedia Tools and Applications, 2015, 2687-2715, ISSN:1573-7721.
[4] L.Z, S.Z, F.J, Y.Q, J.Z, Y.G and H.Z. BoMW: Bag of Manifold Words for One-shot Learning Gesture Recognition from Kinect. IEEE Transactions on Circuits and Systems for Video Technology ( Volume: PP, Issue: 99 ), 2017, Print ISSN: 1051-8215, Electronic ISSN: 1558-2205.
[5] Xianghua Tan, Yueqiang Jin, Guizhen Feng and Xiao Jiang. Gesture Segmentation Based on YCrCb Ellipse Skin Model and Background Subtraction. Proceedings of 2015 IEEE International Conference on Progress in Informatics and Computing(PIC 2015 V1), 2015.
[6] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. International Conference on Computer Vision & Pattern Recognition (CVPR ’05), Jun 2005, San Diego, United States. IEEE Computer Society, 1, pp.886 – 893, 2005.
[7] Overett. Creating robust high-throughput traffic sign detectors using centre-surround HOG statistics. Machine Vision and Applications, 2014, 713-726.
[8] Sujata G.Bhele and Vijay H. Mankar. Recognition of Faces Using Discriminative Features of LBP and HOG Descriptor in Varying Environment. Computational Intelligence and Communication Networks (CICN), 2015 International Conference on, 2015, ISSN: 2472-7555.