Mouse Cursor Control Using Hand Gesture Recognition Based on PHOG-Improved LBP and K-NN Classification

Jin Yi, Kang Chen, Jun Wang, Mingmin Ou and Yi Wu

College of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350108, China.

E-mail:18162350056@163.com

Abstract. A real-time mouse cursor control system using static gestures was designed to achieve the goal of human-computer interaction (HCI) in this paper. The system uses the computer's own camera or external USB camera to collect video data, detect and recognize gestures of people in the video, and control the cursor movement or click in real time based on the gesture. To improve the recognition accuracy of hand gesture images detection, a method of hand gesture images detection based on PHOG + Improved LBP + K-NN was proposed in this paper. In order to improve the real-time performance of the system, the system determines whether the current frame has human hands by skin color detection. When human hands are detected, Pyramid Histogram of Oriented Gradients (PHOG) features and the improved Local Binary Pattern (LBP) features are further extracted. After fusing PHOG features and improved LBP features, k-nearest neighbor classification (K-NN) is used to implement gesture recognition. Six different gestures were tested 50 times with different angles, different lights and no skin tone in the background. The experimental results show that the system has good recognition performance.

1. Introduction
Hand gesture recognition technology is one of the hottest areas of research in the field of HCI. Supported by hand gesture recognition technology, computer devices and robots can be controlled more naturally and efficiently. By getting rid of the limitation of traditional HCI such as keyboard, mouse and remote control, the friendliness of HCI can be improved greatly [1].

Hand gesture recognition has gone through two stages in the development process. The first stage is gesture recognition technology represented by wearable equipment. A lot of sensors are installed in these devices. The second stage is computer vision-based gesture recognition technology. In second stage, electronic devices such as cameras are used to capture, identify and track objects instead of human eyes, and corresponding processing and analysis are carried out to gain the final results[2]. In the computer vision-based gesture recognition technology, the usual method is to extract the gesture features first and then use the classifier for classification and recognition. LBP and HOG are more commonly used gesture feature descriptors. HOG feature has good geometric invariance and optical invariance and is usually used as the feature to extract the edge and contour information of the target image. LBP plays an important role in describing local texture feature information of images. With more and more in-depth research on LBP, its variations become more and more and its applications become more and more extensive[3-5].
Although there have been some researches on gesture recognition, some inherent problems still exist in this field[6]. Studies have demonstrated that hand gestures with different scales, angles and grayscale can reduce the recognition rate [7-8]. In this paper, a real-time mouse cursor control system using static gestures was designed. To break through the limitation of single feature recognition, a method of combining PHOG features with improved LBP features was proposed in this paper to describe gesture features. K-NN was used as a classifier to identify the gesture of the current frame by comparing the distance between the target feature vector in the database and the actual feature vector.

2. Proposed Methodology

2.1. System design

A real-time mouse cursor control system using static gestures was designed to achieve the goal of human-computer interaction (HCI) and a method of hand gesture images detection based on PHOG + Improved LBP + K-NN was proposed in this paper. In this system, six static gestures are selected as the input of the mouse cursor control system. The flowchart of the gesture recognition system proposed in this paper is shown in figure 1. The computer's camera collects video data from users at 25 fps. In order to improve the real-time performance of the system, a frame is grabbed every 5 frames for processing. Each captured frame data needs to be preprocessed first, including RGB to YCbCr and median filtering. After preprocessing, skin color detection based on elliptic model was performed on the preprocessed frame data. If skin color was not detected, the video data of the next frame was processed. If skin color was detected, the gesture region of the current video frame was extracted based on skin color, and the image was normalized to $256 \times 248$. After extracting PHOG features and the improved LBP features and fusing them, the features were compared with those in the feature template database. 50 fusion features for each gesture are stored in the template database. K-NN algorithm is used to obtain recognition results, and the cursor movement is controlled according to the recognition results.

![Flowchart of the gesture recognition system](image-url)
2.2. Skin detection algorithm
The data format obtained from the camera is RGB format. In RGB space, skin color is sensitive to light changes, and the three color components have strong correlation, which is not applicable to skin color model. YCbCr is a discrete color space that separates brightness information from other information, and has a good effect of skin color clustering. This color space decomposes RGB into brightness information and chroma information for processing respectively. The mutual conversion formula of the two spaces is as follow:

\[
\begin{bmatrix}
Y \\
Cb \\
Cr
\end{bmatrix} =
\begin{bmatrix}
16 & 65.481 & 128.553 & 24.966 \\
128 & -37.797 & -74.203 & 112.000 \\
128 & 112.000 & -93.786 & 18.214
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\]

(1)

The commonly used skin color detection methods include threshold definition and elliptical skin color model. Threshold definition is the most commonly used method for skin color detection. For example, the observation of a large number of skin color samples shows that the skin color on plane CbCr is mainly concentrated in the range of \(77 < Cb < 127\) and \(133 < Cr < 173\). The advantage of threshold definition method is simple and fast, but its misjudgment rate is very high, it is easy to mistake some non-skin color areas for skin color areas. In the process of studying YCbCr, R.L.Hsu found through experiments that the skin color distribution on the CbCr plane conforms to the elliptic distribution, and proposed the elliptic clustering model. The model can be expressed in equations (2) and (3) as follows:

\[
\frac{(x - ec_x)^2}{a^2} + \frac{(y - ec_y)^2}{b^2} = 1
\]

(2)

\[
\begin{bmatrix}
x \\
y
\end{bmatrix} =
\begin{bmatrix}
\cos \theta & \sin \theta \\
-\sin \theta & \cos \theta
\end{bmatrix}
\begin{bmatrix}
Cb - cb_0 \\
Cr - cr_0
\end{bmatrix}
\]

(3)

The criteria are as follows:

\[
D(Cb, Cr) = \begin{cases} 
1 & \frac{(x - ec_x)^2}{a^2} + \frac{(y - ec_y)^2}{b^2} \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

(4)

In the formula (3) and (4), \(\theta\) is equal to 2.53, \(c_x\) is equal to 109.38, \(c_y\) is equal to 152.02, \(ec_x\) is equal to 1.60, \(ec_y\) is equal to 2.41, \(a\) is equal to 25.39, \(b\) is equal to 14.03.

2.3. PHOG feature
HOG feature has good geometric invariance and optical invariance and is usually used as the feature to extract the edge and contour feature information of the target image. HOG constructs the feature by calculating and statistical the histogram of the gradient direction in the local area of the image. For an image, the main principle is that the direction of the density gradient of the local object can well describe the appearance features and shape features of the edges. Therefore, the essence of HOG is the statistics of gradient. The whole image needs to be normalized in gamma space to suppress the interference of noises in the process of HOG computing. The Gamma compression formula is as follow:

\[
I(x, y) = I(x, y)^{\gamma_{\text{gamma}}}
\]

(5)

In formula (5), gamma is usually equal to 0.5.

The formulas of horizontal gradient and the vertical gradient of pixel \((x, y)\) in the image are as follows:

\[
\begin{align*}
\nabla_x &= \frac{I(x+1, y) - I(x-1, y)}{2} \\
\nabla_y &= \frac{I(x, y+1) - I(x, y-1)}{2}
\end{align*}
\]
\[ G_x(x, y) = H(x + 1, y) - H(x - 1, y) \]
\[ G_y(x, y) = H(x, y + 1) - H(x, y - 1) \]  

In formula (6), \( H(x, y) \) denotes the pixel value, \( G_x(x, y) \) denotes the horizontal gradient of the pixels \((x, y)\), and \( G_y(x, y) \) denotes the vertical gradient.

The magnitude \( G(x, y) \) and orientation \( \theta(x, y) \) of pixels \((x, y)\) can be formulated as follows:

\[ G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \]
\[ \theta(x, y) = \arctan \left( \frac{G_x(x, y)}{G_y(x, y)} \right) \]  

The calculation of Pyramidal Histogram of Oriented Gradient (PHOG) is based on HOG. Firstly, the HOG feature of all small cells in each block is concatenated as the HOG feature descriptor of block, and then the HOG feature vector of all blocks is cascaded to form the HOG feature descriptor of the layer image. Finally, the HOG feature vector of all layers is cascaded to obtain the PHOG feature descriptor of the image. As a variant of HOG, PHOG is also a descriptor. The performance of PHOG is improved by improving the spatial or three-dimensional information of local area. Therefore, PHOG is used to represent one of gesture features in this paper.

2.4. Improved LBP feature

LBP is one of operator used to describe the local texture feature information of images. The original LBP algorithm compares the gray values of 8 adjacent pixels in a 3 × 3 window with the center pixel of the window as the threshold. In this way, an 8-bit binary vector can be extracted. The LBP feature of the central pixel is obtained by converting the 8-bit binary vector into decimal. The biggest disadvantage of the original LBP operator is that it only covers a small area within a fixed radius and obviously cannot meet the needs of different scale textures. In order to adapt to the texture features of different scales and meet the requirements of gray scale invariance and rotation invariance, Ojala et al. improved the LBP operator and extended the square neighborhood of 3 × 3 to any neighborhood of a circle. The improved LBP operator allowed arbitrary multiple pixel points in the circular neighborhood with radius \( R \). Thus, an LBP operator with \( P \) sampling points in the neighborhood of a circle of radius \( R \) is obtained.

There is too much noise in gesture image under complex environment, which will cause excessive interference factors in the detection process. However, the traditional LBP algorithm has relatively poor anti-noise ability. In order to enhance the anti-noise capability of the system and better describe the local features of gestures, this paper proposed an improved LBP algorithm: MLBP(Mean Local Binary Pattern). The improved algorithm is based on the traditional circular neighborhood operator with radius of \( R \). The maximum and minimum values of the sampling points are removed, and the mean value of the remaining elements is used to replace the original threshold value of the central pixel. The calculation process is as follow:

![Figure 2. 5 × 5 circular neighborhood LBP operator](image-url)
A 5 × 5 pixel template is defined in figure 2, in which the white circle is the central pixel to be calculated and the black circle is the adjacent pixel. In this paper, the coordinates of the central pixel point are assumed to be \((X_c, Y_c)\) , and the coordinates of the adjacent pixel point are assumed to be \((X_p, Y_p)\). The conversion formula between the coordinates of the center pixel point and the adjacent pixel point are as follows:

\[
X_p = X_c + R \cos\left(\frac{2\pi P}{P}\right) \\
Y_p = Y_c - R \sin\left(\frac{2\pi P}{P}\right)
\]  

(8)

The coordinate values calculated by formula (8) are not necessarily integers, so it is necessary to calculate the pixel value of the sampling point by using bivariate interpolation. The calculation formula of bilinear interpolation is as follow:

\[
f(x, y) \approx [1 - x] \begin{bmatrix} f(0, 0) & f(0, 1) \\ f(1, 0) & f(1, 1) \end{bmatrix} [1 - y]
\]  

(9)

After obtaining the pixel value, the LBP characteristic operator is calculated as follow:

\[
LBP_{p,R} = \sum_{i=0}^{P-1} 2^i \cdot S(g_i - g_c)
\]  

(10)

\[
S(x) = \begin{cases} 1, & s.t \quad x \geq 0 \\ 0, & otherwise \end{cases}
\]

In formula (10), \(g_i\) denotes the gray value of the adjacent pixel, and \(g_c\) denotes the gray value of the central pixel. \(P\) denotes the number of pixels in the adjacent of the sampling point, \(R\) denotes the radius of the neighborhood.

The MLBP algorithm proposed in this paper is as follows. In formula (11), \(g_m\) denotes the mean value of the remaining elements after removing the maximum and minimum values of the sampling points.

\[
\left\{h_0, h_1, \ldots h_{P-1}\right\} = \text{sort}\left\{g_0, g_1, \ldots g_{P-1}\right\} \\
g_m = \frac{1}{P - 2} \sum_{i=4}^{P-1} h_i \\
MLBP_{p,R} = \sum_{i=0}^{P-1} 2^i \cdot S(g_i - g_m) \\
S(x) = \begin{cases} 1, & s.t \quad x \geq 0 \\ 0, & otherwise \end{cases}
\]  

(11)

Many binary patterns can be generated by an LBP operator. For LBP operator containing \(P\) sampling points in a circular neighborhood of radius \(R\), \(P^2\) patterns will be generated. Obviously, the number of binary patterns increases dramatically as the number of samples in the neighborhood increases. So many binary patterns are unfavourable to texture extraction and feature classification. In this paper, uniform pattern is applied to reduce the dimension of the LBP operator. The uniform pattern means a pattern in which the number of transitions between 0 and 1 in a binary sequence is less than or equal to 2. The number of patterns is reduced from \(P^2\) to \(P(P - 1) + 2\), where \(P\) denotes the number of sampling points in the neighborhood. For the 8 sampling points in the neighbourhood, the 256 binary patterns will be reduced to 58 when uniform pattern is applied, which makes the dimension of the feature vector smaller and reduces the interference from high-frequency noise.
2.5. K-Nearest Neighbor Classification

K-NN is used as a classifier for gesture feature recognition in this paper. 50 images of each of the six gestures were collected as a sample training set. The PHOG+MLBP feature vectors of the sample training set were extracted and labelled as the template database for K-NN classification. In the gesture recognition process, the Euclidean distance between the target feature vector in the database and the actual feature vector is calculated. The formula for calculating Euclidean distance is as follow:

\[ D = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \]  

(12)

In formula (12), \( P = (p_1, p_2, \cdots, p_n) \) denotes the feature vector of the input image, and \( Q = (q_1, q_2, \cdots, q_n) \) denotes the feature vector of the template database. According to the calculation results, the top K features closest to the template database are selected, and the category with the highest frequency among the K features is the recognition result.

3. Experimental Results and Analysis

The system was implemented using Matlab R2014a Software. Matlab itself cannot realize the simulation of mouse cursor movement and click. The java.awt.robot class is called in MATLAB in this paper to achieve these functions. After referring to the java.awt.Robot class and instantiating it in MATLAB, by setting the mouse cursor position \((x, y)\), we can flexibly control the mouse movement to any position on the screen. Where \((x, y)\) are the coordinates of the mouse cursor relative to the upper left corner of the screen.

Table 1 shows the experimental results of the hand gesture recognition system proposed in this paper. Six different gestures were tested 50 times with different angles, different lights and no skin tone in the background. Because the classifier chooses K-NN algorithm, the selection of K value is very important. After k-fold cross-validation, \( k = 3 \) is finally selected in this paper. Too little data volume of the training template cannot guarantee the recognition accuracy of the system, and too much will reduce the operating efficiency of the system. In order to improve the accuracy of system recognition and the efficiency of system operation, the data amount of the training template should be appropriately selected. This paper selected 50 images of each gesture as database templates.

| Gesture        | No. Person | Images properly recognized | Accuracy |
|----------------|------------|----------------------------|----------|
| One Finger     | 50         | 46                         | 92%      |
| Two Finger     | 50         | 44                         | 88%      |
| Three Finger   | 50         | 43                         | 86%      |
| Four Finger    | 50         | 40                         | 80%      |
| Open hand      | 50         | 45                         | 90%      |
| Fist           | 50         | 40                         | 80%      |
| Total          | 300        | 258                        | 86%      |

4. Conclusion

The proposed system can control the cursor movement by using static hand gestures without other traditional hardware input. Computer's camera is called in MATLAB to input the gesture information. The PHOG feature and the improved LBP feature of the gesture image are extracted, and the K-NN algorithm is used to classify it into one of the six gestures. The recognition accuracy of the proposed system is about 80%. This paper only makes a preliminary exploration of gesture recognition.
technology and its application. It is believed that in the near future, more and more HCI devices based on hand gesture recognition will appear, and human social life will be greatly enriched.

References
[1] Escalera, Sergio, Isabelle Guyon, and Vassilis Athitsos, eds. Gesture recognition[M]. Berlin, German: Springer, 2017:2-26
[2] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]// Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005:886-893
[3] J. Sun, X. Wu, “Infrared target recognition based on improved joint local ternary pattern,” Optical Engineering, vol. 55, no. 5, 2016
[4] A. Asthana, S. Lucey, and R. Goecke, “Regression based automatic face annotation for deformable model building,” Pattern Recognition, vol. 44, no. 10, pp.2598-2613, 2011
[5] H. Lategahu, S. Gross, and T. Stehle, et al, “Texture classification by modeling joint distributions of local patterns with Gaussian mixtures,” IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, vol. 19, no. 6, pp.1548-1557, 2010
[6] P. K. Pisharady, M. Saerbeck, “Recent methods and databases in vision-based hand gesture recognition: A review,” Computer Vision and Image Understanding, vol. 141, pp.152-165, 2015
[7] B. K. Chakraborty, D. Sarma, and M. K. Bhuyan, “Review of constraints on vision-based gesture recognition for human-computer interaction,” IET Computer Vision, vol. 12, no. 1, pp.3-15, 2018
[8] C. Zou, Y. Liu, J. Wang, et al, “Deformable part model based hand detection against complex backgrounds,” Advances in Image and Graphics Technologies, IGTA 2016 Book Series: Communications in Computer and Information Science, vol 634, pp.149-159,2016