Monocular Multi-feature Fusion Hand Gesture Recognition Method Based on Multi-core Learning

Kai Zhuang
School of Information science and engineering, Yanshan University, Qinhuangdao 066004, China.
Email: zhuangkai1998@outlook.com

Abstract. Based on multi-core learning, this paper uses the method of skin color segmentation, fingertip distance detection and extraction and gesture contour extraction, which realizes the fusion of local features and overall features and makes the accuracy of hand gesture classification, reach 92%. Compared with traditional gesture recognition based on binocular camera, our algorithm using multi-core learning has the same accuracy as the recognition based on binocular camera, and the complexity and cost of our algorithm are lower. Thus, our algorithm has broader application prospects. Compared with gesture recognition based on monocular camera, our algorithm is more suitable for the background environment of gesture classification. In other words, it has better generalization ability.

1. Introduction
Hands are one of the most important tools of human beings which promote the development of human civilization [1]. With the rapid development of computer technologies, the interactive application between mechanical equipment and humans is exponentially increasing and the novel interactive technologies between humans and computers conforming to the habit of interpersonal communication has become a research hotspot. From the initial technology limited to speech recognition and control, it has evolved into tracking of motion, location and hand gestures [2]. According to the difference of application purposes, hand gestures can be divided into control gestures, conversation gestures, communication gestures and operation gestures.

Hand gestures recognition is a technology utilizing mathematics algorithm containing computer graphics, supported by camera, data glove and other input tools. As to the collected information, such as the orientation and angle of the palms and fingers, the technology makes judges, analysis and correct responses [3]. At the time being, hand gesture recognition has already had widespread applications, for instance, in terms of human-computer interaction, gestures are used as input to control the operation of the machine so that a richer bridge between machine and humans than the original text user interface or even the GUI (graphical user interface). Moreover, hand gesture recognition is vital for the progress of other technologies. Taking behavior analysis as an example, it comprehends people’s posture, hand gestures and so on to judge whether a person's behavior is abnormal behavior, which can effectively combat criminal behavior.
The initial gesture recognition mainly uses the direct detection of the machine equipment to obtain the spatial information of the hand and each joint. The typical representative equipment is data gloves and so on. However, the data glove is limited by the naturalness of the gesture and few recognizable hand gestures. The emergence of optical marking method has replaced the data glove, which can provide better recognition result, but the equipment is complicated and it is difficult to be applied for. The accuracy and stability of the wearable device directly detecting gestures is higher, but it limits the natural expression of gestures. Compared to the recognition system of wearable devices, a vision-based gesture recognition system that utilizes gesture image information captured by a video capture device to process gestures by computer vision technology is more convenient for operators to have interaction between humans and machine. This method doesn’t need complicated and expensive equipment and is handier. Thus, in the natural scene, vision-based hand gesture recognition technology will become the development trend in the future.

Vision-based 3D hand gesture recognition has a fast progress in terms of accuracy and dataset quality [4]. However, it mostly utilizes depth camera or multi-view devices which are expensive and have poor versatility so that it leads to the lack of large-scale application in actual scenarios. In contrast, single-view camera is cheap and practical. Therefore, research on 3D gesture recognition based on single-view camera has become a hot spot.

A system based on a monocular gesture recognition usually models human hand as a pixel or pixel block [5]. It utilizes the apparent features such as skin color, contours and so on to analyze the motion parameters of the gesture as a whole. Du et al. realize a simple gesture recognition system by a single camera through skin color segmentation, feature extraction and finite state machine training and recognition. However, the method is harsh on the conditions of the use environment, and the uncertainty of skin color detection is large.[6]Utilizing a fixed camera realizes the recognition of fingertip orientation, but the accuracy of recognition is seriously affected by the location of the testers. Kolesnik et al. using a single camera on the ceiling and through the identification of the direction of the fingertip, realize the movement of objects in the virtual environment which is controlled by a single gesture. However, only when the orientation of movement of fingers is horizontal can it be recognized.

Nowadays, on one hand, due to the high degree of freedom of hand gestures, the deformation is difficult to control. Under the factors of position dynamics and rotation changes, the recognition of gestures is seriously disturbed and the complexity is large. On the other hand, the problem of illumination intensity and background dynamics in the actual scene brings great challenges to the segmentation of gestures, directly affecting the extraction and recognition of gesture features. As to the problems above, we propose a monocular multi-feature fusion gesture recognition method which is based on multi-core learning. By analyzing multiple feature information of a gesture and utilizing multi-core learning methods, it can build a multi-feature model that realizes effective fusion of multiple hand gesture features and improves the accuracy and robustness of hand gesture recognition. The simulation results show that the proposed algorithm is optimized in terms of accuracy and robustness.

2. Basic Theory of Single-view Hand Gesture Segmentation

Hand gesture recognition is mainly divided into three processes: gesture segmentation, feature extraction and recognition. Gesture segmentation is the basis of the whole recognition process and feature extraction is the core of recognition. The accuracy of gesture segmentation will seriously affect the accuracy and real-time of the latter recognition part. This paper comprehensively utilizes multiple feature fusion segmentation methods to construct the local description features and overall description features of gestures, and achieve effective classification of hand gestures.

2.1. Skin Color Segmentation Model

The skin color segmentation model mainly uses the spatial clustering feature of skin color on the color space, to separate the region of interest from the background environment. Due to the different clustering characteristics of skin color in different color spaces, choosing the right color space is more beneficial
for clustering of skin colors. It has been shown that in the YCrCb space, the skin color is less affected by the illumination brightness and the clustering characteristics are better. Therefore, the YCrCb space is used as the color space of the gesture segmentation in the skin color detection model. We convert RGB space to YCrCb space using equation (1):

\[
\begin{align*}
Y &= 0.257 \times R + 0.564 \times G + 0.098 \times B + 16 \\
Cb &= -0.148 \times R - 0.291 \times G + 0.439 \times B + 128 \\
Cr &= 0.439 \times R - 0.368 \times G - 0.071 \times B + 128
\end{align*}
\]

(1)

2.2. Hand Gesture Overall Contour Extraction

Kass et al. proposed the Snake dynamic contour model in 1987, which can segment the target from the complex background, effectively tracking the deformation of the target, and is not sensitive to noise and contrast. In this paper, when the overall contour of the gesture is extracted, a continuous smooth closed contour curve is established through the Snake model, and the elastic deformation of the gesture is combined with the local features of the image to achieve image segmentation. The basic Snake model is as in equation (2):

\[
l(s) = [x(s), y(s)] \quad s \in [0,1]
\]

(2)

The Snake model consists of straight lines connected by these points, where \(x(s), y(s)\) are the coordinate positions in the image, and \(s\) is the independent variable describing the boundary of the contour. Therefore, the energy function of Snake model is:

\[
E = \int_0^1 \frac{1}{2} [\alpha(s)|l(s)'|^2 + \beta(s)|l(s)''|^2] + \frac{1}{2} E_{ext}[l(s)] ds
\]

(3)

The first control contour in equation (3) is a smooth curved surface, the second driving contour is a smooth curve, and the third term is external energy, making contour close to the high gradient position of the actual image. \(\alpha, \beta\) are parameters that control the elasticity and bending degree of the contour. The initial contour of the gesture can be constructed by the Snake algorithm, as shown in Figure 1.

![Figure 1](image)

**Figure 1.** Initial hand contour consisting of 200 control points

2.3. Hand Gesture Local Feature Extraction

In order to better build a dynamic model of hand gesture, we choose to extract local feature, namely hand joints, shown in Figure 2. Inspired by recent work on hand joint prediction, we train a CNN, the RegNet, which is derived from the ResNet50 architecture, that predicts 2D and 3D positions of 21 hand points. The RegNet is based on a residual network consisting of 10 residual blocks, and its architecture is shown in Figure 3.
After operating the network, we get the 2D joint predictions in the form of heatmaps in image space and the 3D joint coordinates relative to the root joint. Those are shown in figure 4.

**Figure 4.** 2D Joint Heatmaps and 3D Joint Locations

Therefore, we are able to compute the length of five fingers, distance between adjacent finger tips and distance from finger tips to palm because we get the 3D coordinates of the 21 hand joints. We define the length of five fingers as

\[ L_i = \sqrt{(l_{i+15,x} - l_{i,x})^2 + (l_{i+15,y} - l_{i,y})^2 + (l_{i+15,z} - l_{i,z})^2} \quad i \in [1,5] \]  

(4)

\( L_i \) represents the length of the \( i \)-th finger, and the parameter \( l_{i,x,y,z} \) means the 3D coordinate of the \( i \)-th joint. We also define the distance between adjacent finger tips as

\[ D_i = \sqrt{(l_{i+16,x} - l_{i+15,x})^2 + (l_{i+16,y} - l_{i+15,y})^2 + (l_{i+16,z} - l_{i+15,z})^2} \quad i \in [1,4] \]  

(5)

\( D_i \) represents the distance between \( i \)-th finger and \( i + 1 \)-th finger. The distance from finger tips to palm is also defined as
\[ E_i = \sqrt{(l_{i+15,x} - l_{21,x})^2 + (l_{i+15,y} - l_{21,y})^2 + (l_{i+15,z} - l_{21,z})^2} \quad i \in [1, 5] \quad (6) \]

\( E_i \) represents the distance between \( i \)-th finger and palm.

3. Multi-core Learning

3.1. Multi-feature Fusion Method Based on Multi-core Learning

In the \( \mathcal{Z} \) space, for any sample \( x \in \mathcal{X} \), the discriminant function in the feature space is:

\[ f(x) = (w, \phi(x)) + b \quad (7) \]

\( \phi(x) \in R^D \) is a sample in the feature space, \( w \) is a weight vector, and \( b \in R \) is a bias. In space \( \mathcal{Z} \), \( f(x) \) is a linear function of \( w \) and \( b \), but nonlinear from the \( \mathcal{X} \) space. The classification decision condition is: if \( f(x) > 0 \), it is determined that \( x \) belongs to the positive class, otherwise it belongs to the negative class. Hence, the minimization problem of single-core SVM problem is:

\[
\begin{align*}
\min_{w,b,i} & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} & \quad y_i f(x_i) \geq 1 - \xi_i, \xi_i \geq 0; \quad i = 1, \ldots, n
\end{align*}
\]

\( \xi \) is a slack variable, and regular parameter \( C > 0 \) is the balance model complexity and empirical risk. When the number of samples is large and the data dimensions are high and irregular, using single kernel function is not able to obtain ideal results. In order to feasibly choose suitable kernel function, researchers proposed the method of using combined kernel function instead of single kernel function. The most basic multi-core function is a set of kernel functions for convex linear combination:

\[ K(x_i, x_j) = \sum_{m=1}^{M} w_m K_m(x_i, x_j) \quad \text{s.t.} \quad w_m \geq 0 \quad (9) \]

\( M \) represents the number of the kernel functions, \( w_m \) represents the weight of \( m \)-th kernel function and \( K_m \) represents \( m \)-th basic kernel function.

It is difficult to select the appropriate kernel parameter for the learning method of single core multi-features in gesture recognition. In this paper, we use the multi-core learning method to construct the interval maximization model and find the linear/nonlinear combination of the given basic kernel function, which is more flexible than the single-core learning in solving the similarity of the data source [7]. Hence, we utilize multi-core learning to fuse overall and local features, namely hand contour features and hand joints features. In this way, we are able to determine their respective kernel functions. Finally, a hand classifier will be created.

3.2. Gesture Recognition Method Based on Multi-core Optimal Combination Learning

When selecting the support vector machine kernel function, there is no unified method to choose the kernel function [8]. So most researchers use the empirical method, which leads to the selection of the kernel function and its parameters has a great influence on the recognition result. Therefore, the support vector machine classification algorithm based on single kernel function is difficult to meet the needs of complex classification problems [9]. Especially for multi-source heterogeneous data classification, the classification accuracy of single-core algorithm is even worse. This paper integrates multi-gesture features through multi-core learning methods to improve the generalization ability of support vector machines. In the process of recognizing gestures, this paper uses the texture, overall contour and local features of the fingertips to unify the three heterogeneous features into the same
decision structure through the basic kernel, and uses the multi-core learning theory to make three weighted features into a fusion core. At present, there are many types of basic kernel functions. In practical applications, a kernel function is effective for classifying some information samples, but it is not effective for other samples. Therefore, different kernel functions need to be selected based on analysis of different information.

Since the number of skin texture features is large, but the recognition classification is relatively simple, this paper selects a local binary pattern (LBP) kernel function for the skin texture feature and extracts the skin texture feature from the brightness change of the local pixel of the gesture. After convoluting the corresponding gray value of the gesture image, we can get the transformed value:

$$LBP(c) = \sum_{i=0}^{7} s(p_i, c)2^i$$

$c$ represents pixel $(x_c, y_c)$, and $p$ represents the gray value of 8 neighborhoods of pixel $c$. In order to perform gesture image recognition, this paper uses histogram as a measure of similarity:

$$d(F_1, F_2) = \frac{\sum_{i=0}^{M} \min(H_i^{F_1}, H_i^{F_2})}{W \times H}$$

$M$ is the total number of the image blocks. $W$ and $H$ are normalized image widths and heights.

The gesture image is converted into the YCbCr model. Besides, the luminance information is separated by $Y$, and the color information is stored by using $Cb$ and $Cr$, which represents color difference. Figure 5 shows the gesture skin color segmentation results of the LBP operator in the YCbCr space.

![Figure 5. LBP operator’s gesture skin color segmentation](image)

For the local fingertip distance feature, the radial basis RBF kernel function is adopted [10]. Because of its simple model and relatively few parameters, it can convert nonlinear problems into linear problems. The RBF kernel function can be expressed as:

$$\kappa(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2 / \sigma^2}, \gamma > 0$$

$\sigma$ represents the range of the action of radial basis kernel function. $x_j$ represents the center of the function. The optimal classification decision is:

$$f(x) = \text{sgn}\left\{\sum_{j=1}^{n} \alpha_j^* y_j \exp\left\{-\frac{\|x-x_j\|^2}{\sigma^2}\right\} + b^*\right\}$$

The advantage of the radial basis function is that the number of training samples in the feature space does not affect the classification result.
Figure 6. distance detection of hand area

The integrity and continuity of gesture contour extraction directly affect the image quality of gesture recognition. Although existing methods such as Sobel operator and canny operator are simple and efficient, the extracted contours are incomplete, have breakpoints, and have redundant background information. In this paper, the Gaussian mixture model is combined with the snake operator to remove the background noise information and make the gesture contour more continuous. Hence, a more complete gesture contour can be extracted. The probability density function of the mixed Gaussian model can be expressed as:

$$P(x_i) = \sum_{j=1}^{M} \alpha_j N_j(x_i; \mu_j, C_j)$$

(14)

$M$ is the number of single Gaussian models and $\alpha_j$ is the weighting factor of each Gaussian model. The GMM algorithm can describe the feature of each pixel of the image, and update the background model of the image captured by the camera, and compare it with the current image pixel to obtain a denoised foreground image. We combine GMM model with equation (3) to construct the contour connection model:

$$k = \frac{y_2 - y_1}{x_1 - x_2}, x_1 \neq x_2$$

(15)

$(x_1, y_1)$ is the end point of the track of the gesture outline, and $(x_2, y_2)$ is the pixel coordinates of the previous point. We achieve the consistence of gesture contours through slope smoothing.

Figure 7. Contour detection of hand area

We get the overall architecture diagram of multi-core learning for gesture recognition.
4. Experiment Results
In this paper, 8 sets of gesture data are collected. As shown in Figure 9, for each set of gestures, the skin color information, fingertip distance information and gesture contour information are detected by multi-core learning method as the main features of the gesture.

The following results are obtained by performing skin color segmentation, fingertip distance measurement, gesture contour extraction for each set of gestures, which are shown in figures 10, 11 and 12.
Figure 10. Gesture skin color segmentation

Figure 11. Extraction of gesture fingertip distance information

Figure 12. Extraction of gesture contour
For the above eight sets of gestures, under the same condition of background environment, illumination, interference and so on, we select the continuous 100 frames to test the accuracy of the algorithm proposed in this paper. The results are shown in the following table:

| Method of proposed | Accurate detection | Not detected | False detection number of images | Accuracy |
|-------------------|--------------------|--------------|----------------------------------|----------|
| 92                | 5                  | 3            | 100                              | 92%      |

Through the actual gesture detection experiment, the single-core weights in the multi-core are re-adjusted to suit different gesture templates. Finally, we obtain the optimal weight ratios are: skin color detection takes up 16%, fingertip distance detection takes up 64%, and contour detection takes up 20%.

The accuracy of ours is equal to the traditional binocular camera-based gesture recognition accuracy. Moreover, we has lower complexity than current gesture recognition algorithms, so we can run faster than theirs. Despite what we have done, our algorithm is only applicable to static gesture recognition rather than dynamic gesture recognition. In other words, we can’t recognize some dynamic gestures, such as waving hands, clapping hands and so on.

To address this, we will focus on dynamic gesture recognition in the future research and try to achieve real-time recognition.

5. Conclusion

We propose a multi-feature fusion recognition method based on multi-core learning for the problem of low accuracy and difficulty of monocular gesture recognition. Through the methods of gesture skin color detection and segmentation, fingertip distance detection and extraction, and gesture contour characterization, we realize the fusion of local features of gestures and overall feature information, which effectively solves the problem that traditional single-core gesture learning algorithms are difficult to meet complex classification. Using multi-core learning method can better adapt to the background environment of gesture classification, and it can improve the generalization ability of monocular camera gesture recognition. The experimental simulation proves that the accuracy of gesture recognition is 92%, which is equivalent to the traditional binocular camera-based gesture recognition accuracy. However, the algorithm has low complexity, low cost and better application prospects.

6. Reference

[1] Pavlovic V I, Sharma R, Huang T S. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review [J]. IEEE Trans Pattern Analysis & Machine Intelligence, 1997, 19(7):677-695.
[2] Yang M H, Ahuja N, Tabb M. Extraction of 2D Motion Trajectories and Its Application to Hand Gesture Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(8):1061-1074.
[3]Wachs J P, Kölsh M, Stern H, et al. Vision-based hand-gesture applications[J]. Communications of the Acm, 2011, 54(2):60-71.
[4]Laskar M A, Das A J, Talukdar A K, et al. Stereo Vision-based Hand Gesture Recognition under 3D Environment[J]. Procedia Computer Science, 2015, 58:194-201.
[5] Chonbodeechalermroong A, Chalidabhongse T H. Dynamic contour matching for hand gesture recognition from monocular image[C]// International Joint Conference on Computer Science & Software Engineering. 2015.

[6] Lee L K, An S Y, Oh S Y. Robust fingertip extraction with improved skin color segmentation for finger gesture recognition in Human-robot interaction[C]// Evolutionary Computation. IEEE, 2012.

[7] Wang D, Chaudhari N S. A multi-core learning algorithm for Boolean neural networks[C]// International Joint Conference on Neural Networks. 2003.

[8] Yan-Ling L, Lei L I, Zhou M M, et al. A NEW FUZZY SUPPORT VECTOR MACHINE BASED ON MIXED KERNEL FUNCTION[C]// International Conference on Machine Learning & Cybernetics. 2009.

[9] Chen B, Duan L, Hu J. Composite kernel based SVM for hierarchical multi-label gene function classification[C]// International Joint Conference on Neural Networks. 2012.

[10] Le D, Mizukawa M. Fast fingertips positioning based on distance-based feature pixels[C]// Third International Conference on Communications & Electronics. 2010.