Vision-based hand gesture recognition system

Shun-Feng Su, Chung-Ju Liao, Ming-Chang Chen
Department of Electrical Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan
sfusa@ntust.edu.tw, m9707310@mail.ntust.edu.tw, maximchen.chen@gmail.com

Hsun Li
Department of Information Technology
Lee-Ming Institute of Technology,
New Taipei City, Taiwan
E-mail: ihsumlee@gmail.com

Chen-Chia Chuang
Department of Electrical Engineering
National Ilan University
cechuang@niu.edu.tw

Abstract — The paper proposes a hand gesture recognition system based on vision technology. Unlike other studies that focus on the stage of hand gesture recognition, many problems are considered, such as detection of when a hand reaches the field of the camera view or detection of a full palm. There are four stages in the proposed system: detection of the appearance of hands, segmentation of hand regions, detection of full palm, and hand gesture recognition. Detection of the appearance of hands is to find out when a hand appears in the front of the camera. Moreover, some morphological techniques, along with two stage skin color detection, are used to alleviate the effect of noise. The two stage skin color detection approach is adopted from the idea of handling outliers to extract the palm from complicated backgrounds. Following that, detection of full palm is conducted to know whether the hand reaches beyond the field of the camera view. The concept of ergonomics is employed to determine whether the hand is beyond the field of the camera view. Finally, experimental results show that the proposed system is quite promising.

Keywords — Hand gesture recognition; skin color detection; morphological techniques;

I. INTRODUCTION

Nowadays, the development of modern technologies is rapid and has provided much convenient and secure life for humans. Due to the wide use of computers in all kinds of equipment, ubiquitous computing becomes an important idea in the current research. Ubiquitous computing brings technologies to everywhere in humans' society, especially humans' home. In fact, the emergence of ubiquitous computing has been the main thrust of the development of smart home technologies. Due to the capability of ubiquitous computing, smart home has been considered as a possible solution for the situation when people with disabilities may stay home along. Imagine a scenario that in a smart home environment, users can make hand gestures [1] to remotely control devices such as TVs, air conditioners, doors, curtains, etc. In such a situation, users can command devices without contacting them, which is especially helpful for those with physical disabilities. The aforementioned reasons inspire us to develop a hand gesture recognition system that can be integrated as part of a smart home environment.

Due to its closeness to human nature and possible applications in human-computer interaction (HCI), hand gesture recognition has been attracting considerable researcher to focus on this topic in recent years [2-10]. In the literature, two categories of approaches are considered to model hand gestures. The first approach is based on gloves and is to use sensors such as data gloves to measure the position, joint angles and orientation of hands. The second approach is of non-contact type and is to use vision-based techniques to find the position and gesture of hands [11].

The hand gesture recognition system proposed in this study is aimed at being used for the situation that no accompany at home but a mini TV is hung on the collar of the user. Traditional hand gesture studies are to consider hand gestures appearing in a simple background. Thus, it is easy to extract those hand gestures from such an image and then to recognize those gestures. It can be expected that in our study those hand gesture images are obtained from complicated background instead of some controlled simple background. Thus, in this study, the system needs to find ways of extracting those hand gestures from a complicated background. A new idea called 2 stage skin color detection is proposed to extract the palm from complicated backgrounds. 2 stage skin color detection adopts the concept of center, a term often used in the field of statistics.

Unlike most studies that focus on the stage of hand gesture recognition, many other problems have to be dealt with in this study, such as detection of when a hand reaches in the field of the camera view or detection of full palm. The concept of ergonomics is employed to divide edges of the frame into four line segments. With these four line segments and frame division, which are an idea of partitioning the frame into inner and outer frames, are proposed in this study to decide whether full palm area is wholly included in the image. The experimental results show that our proposed method is quite promising in the stage of detection of full palm.

II. HAND GESTURE RECOGNITION TECHNIQUES

In vision-based human-computer interface systems, the segmentation of foreground objects such as hands, faces, and so on from the background is a major issue for this study. Three methods are mainly used in segmenting target objects from the background: background subtraction, appearance-based method, and color-based detection. These approaches are introduced in the following, and their performances in our system are also compared and discussed.
The traditional background subtraction requires a static camera to capture an image in which without any target objects in the reference frame. The method can extract foreground objects easily by comparing the difference between the reference frame and the current frame [12]. The problem of using the background subtraction approach is that it needs a static background or a static camera. Nevertheless, the users may move their bodies occasionally or when raising their hands even if they try to keep their body stable may still unwittingly move their bodies.

The second method is called the appearance-based method, which is not easily affected by the change in the environment [13]. The method extracts foreground objects based on their shapes. Nevertheless, the foreground object in this study is a hand, which is deformable, non-rigid and has at least 5 degrees of freedom [14]. Hence, it can be expected that the appearance-based method can neither be applied in this proposed system.

The third method is to extract foreground objects from the background image based on color. This method segments foreground objects only by their color information without considering their shapes. Considering the scenario, the best suited approach for our system is the third one: color-based detection. The selection of color space can be regarded as a primary step in skin color detection. In the next sections, several color spaces often used to detect skin color are introduced. Also, related approaches by using different color spaces are implemented in our study.

III. DETECTION OF THE APPEARANCE OF HANDS

An M×N digital image can be defined as a two-dimensional function in which x and y are the spatial coordinates, and the amplitude, I(x, y), at any pair of coordinates is the gray level of the image at the point. Hence, I(x, y) is used to represent an original input image in the surveillance video. Let n_i be the total number of gray level i in an image. Then P_i, which represents the probability of gray level i, can be obtained as

$$P_i = \frac{n_i}{MN}.$$  

(1)

If an image has L gray levels, the entropy \(H_f\) of the frame is defined as

$$H_f = -\sum_{i=1}^{L} P_i \log P_i.$$  

(2)

A. Segmentation of Hand Regions

In our system, the user may wear wrist artifacts or rings, and skin-color-like noises are allowed to exist in the background as long as they are smaller than the hand. Hence, background noises often exist after the skin color pixels of input images are detected as shown in Fig. 1. In this section, further steps shall be implemented to remove noises, to reconnect fingers to the palm and to extract hand regions. Segmentation of hand regions is achieved by the following steps. First, skin color detected images obtained from Section II are converted into binary images. Next, the opening operation is used to initially remove possible background noises. In addition, the problem of having rings on fingers can also be resolved in this step by applying the closing operation. Unlike other traditional approaches, both component labeling and the concept of seam line are employed to search for the palm rather than using component labeling alone. Then, the 2nd stage of skin color detection is applied to refine the binary image obtained from the previous step. Finally, some morphological techniques that are implemented in the second step, along with region filling, are used again to get a preferable result. Flow chart of segmentation of hand regions is shown in Fig. 2.

Figure 1. Images after “Detection of the appearance of hands.”

Figure 2. Flow chart of segmentation of hand regions.
B. Detection of Full Palm

Since the mini DV is hung on the middle of the user's collar, he/she is not aware whether the full palm is in the field of view or not. Hence, it is necessary that the system must be able to know whether the current image includes the full palm area or not. Ergonomically speaking, the user will "raise" his/her hand into the camera view. Thus, in this study, we assume the user's arm will move from the edges of the bottom part of the image no matter whether the right or left hand is considered as shown in Fig. 3. Thus, the considered edges are from the middle of the right-hand side downward to the middle of the left-hand side as indicated by thick black lines in Fig. 3.

![Image](image1)

Figure 3. Users reaching their hands

In addition, while the palm in an image is closer to the camera, the size of the palm may exceed the threshold value of a human hand even when the palm is not completely within the field of the camera view. Hence, a concept of dividing the image into two parts called frame division is proposed. In this approach, we consider when the palm is within the center area of the image, it is viewed that the full palm is in the image. To implement this idea, we need to define where the palm is located and where the center area of the image is. For the palm location, we simply consider the centroid of the obtained skin-color object. The centroid \((\bar{x}, \bar{y})\) of a segmented hand region is calculated as

\[
\bar{x} = \frac{\sum_{i=1}^{k} x_i}{k} \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^{k} y_i}{k}
\]

where \(x_i\) and \(y_i\) are \(x\) and \(y\) coordinates of the \(i^{th}\) pixel in the hand region, \(k\) denotes the number of pixels in the region.

C. Hand Gesture Recognition

After the palm region is segmented from the background and filtered by the algorithm proposed, the palm is obtained and then the system needs to know how many fingers have been raised. In order to detect the number of fingers, the hand of the binary image will be transformed into a polar image for recognition [15]. Polar hand image is a popular method of recognizing hand gestures due to its scale, translation and rotation invariance. An example of a polar hand image is shown in Fig. 4. We shall introduce the concept in the following. The centroid of the foreground object is obtained using (3), and the boundary of the foreground object is acquired [16] as

\[
B = (A) = A - (A \Theta B),
\]

where \(A\) is the concerned image and \(B\) is a 3x3 structuring element as \([1 1 1; 1 1 1; 1 1 1]\). \(\Theta\) is an erosion operation. The idea of (4) is to erode the concerned image by a 3x3 structuring element and then subtract it by the original image. Then the distance of each point on the contour from the centroid is calculated. The result of transforming a hand image into a polar image is shown in Fig. 4(b).

The idea is to count how many peaks in the polar image. However, as shown in Fig. 4(b), if the threshold is too low, some unwanted peaks will be included. On the other hand, if the threshold is too high, some peaks will be lost. In order to determine the number of raised fingers, a proper threshold value must be set. A formula is proposed to decide the threshold value of counting peaks in the polar hand images as

\[
\text{heightTh} = (R_{\text{max}} - R_{\text{min}}) \times 0.6 + R_{\text{min}},
\]

where \(R_{\text{max}}\) is the maximum value and \(R_{\text{min}}\) is the minimum value of the polar image.

Nevertheless, there is another problem that the arm region may also be included in the peak numbers as shown in Fig. 5(b). It can be expected that the cross section of an arm is much wider than that of fingers. Hence, in this study, a peak is considered to be a finger only when the width of the peak is within the following interval:

\[
10 < \text{Width} < 35.
\]

With this decision, the arm will not be counted as a finger as shown in Fig. 5(c). Since a hand may show up in the field of view from various angles, the first thing to do is to obtain the angle of the hand. This angle \(\theta\) is obtained as

\[
\theta = \tan^{-1}\left(\frac{x_c - x_m}{y_c - y_m}\right),
\]

where \(x_c\) and \(y_c\) are the coordinates of the centroid, \(x_m\) and \(y_m\) are the coordinates of the middle of foreground pixels along the edges.
IV. EXPERIMENTAL RESULTS

In this session, the experimental results of using the proposed system are presented. The resolution of 320x240 is implemented, and the frame rate is 20 FPS. The system is developed under Microsoft Windows 7. The mini DV used in this study is shown in Fig. 6. The default resolution of the mini DV is 640x480. While the resolution of 640x480 provides higher quality images, it costs more time to process images. Therefore, the resolution of 320x240 is implemented in our experiment for better efficiency in image processing. Table I shows what the characters below each picture represent. The experimental results are shown in Table II. The system judges that the user commands “4, 4, 4, 4, 4,” and all the images are correctly identified. Experimental results are shown in Fig. 7.
V. CONCLUSIONS

This system is operated by mounting a mini DV on the middle of users’ collar, and users are allowed to show up their hands in the field of the camera view. No matter they are moving or at a standstill, wearing long sleeved clothes or short sleeves clothes. The proposed procedure is as follows. First, the number of color pixels is used to decide whether the hand shows up in the field of the camera view. Then, various techniques of segmenting the hand image from the complicated backgrounds are implemented. Following that, the position of the camera inspires us to use ergonomics concept to deal with the problem that hands often reach beyond the image. We cut the edges of the frame into four line segments, along with frame division, to judge whether the image is just partially palm included. Finally, experimental results demonstrate the effectiveness of our proposed scheme for recognizing.

ACKNOWLEDGMENT

This work was supported by the National Science Council and Ministry of Science and Technology, Taiwan, under Grant NSC 101-221E-011-077-MY3 and MOST 103-2628-E-011-020, MOST 103-2221-E-234-004-8, and MOST 103-2221-E-197-017-9.

REFERENCES

<table>
<thead>
<tr>
<th>#</th>
<th>URL/Reference</th>
<th>Words</th>
<th>% Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Internet</td>
<td>72</td>
<td>2%</td>
</tr>
<tr>
<td>4</td>
<td>Internet</td>
<td>28</td>
<td>1%</td>
</tr>
<tr>
<td>6</td>
<td>Internet</td>
<td>12</td>
<td>&lt; 1%</td>
</tr>
<tr>
<td></td>
<td>Author(s)</td>
<td>Title</td>
<td>Conference/Proceedings</td>
</tr>
<tr>
<td>---</td>
<td>---------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>9</td>
<td>vipl.ict.ac.cn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td><a href="http://www.pa.icar.cnr.it">www.pa.icar.cnr.it</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td><a href="http://www.docstoc.com">www.docstoc.com</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Licsar, A.</td>
<td>&quot;User-adaptive hand gesture recognition system with interactive training&quot;</td>
<td>Image and Vision Computing, 20051101</td>
</tr>
<tr>
<td>16</td>
<td>Yanghua Wang.</td>
<td>&quot;Multiple attenuation: coping with the spatial truncation effect in the Radon transform&quot;</td>
<td></td>
</tr>
</tbody>
</table>
domain", Geophysical Prospecting, 1/2003

17 www.inf.ethz.ch
Internet
8 words — < 1%

CrossCheck
8 words — < 1%

CrossCheck
8 words — < 1%

20 "IEEE-ICDCS conference proceeding", 2012 International Conference on Devices Circuits and Systems (ICDCS), 03/2012
CrossCheck
6 words — < 1%

21 Chang, C.-C.. "Feature alignment approach for hand posture recognition based on curvature scale space", Neurocomputing, 200806
CrossCheck
6 words — < 1%

CrossCheck
6 words — < 1%