Contour Based Hand Gesture Recognition Using Depth Data

Le Tran Nguyen, Cong Do Thanh, Tung Nguyen Ba, Cuong Ta Viet, Ha Le Thanh

Human-Machine Interaction Laboratory,
University of Engineering and Technology,
Hanoi, Vietnam
{letn.mi12, congdt_54, tungnb_54, cuongtv, ltha}@vnu.edu.vn

Abstract. This paper presents a contour based method for recognizing hand gesture using depth image data. In this research, a motion based algorithm is used to detect and track human hand. Then, the hand contours are extracted and described by rotation- and scale-invariant feature vectors. Finally, a logistic regression classifier is employed for hand gesture recognition. For evaluating the real-time performance of this hand recognition method, a test-bed application is built. The experimental results exhibit the high recognition accuracy and efficiency of our approach.

Keywords: hand gesture; gesture recognition; contours; depth images.

1 Introduction

Recently, hand gesture recognition has become essential in the field of Human computer interaction. There are countless researches have been focusing on this advanced topic to create natural user interface and improve user experiences by using simple and intuitive hand gestures for free-hand controller [1]. As a result, new problems are emerging such as how to detect the hands as well as tracking and segmenting them from the background in an image sequence. Various methods are discovered and being studied in order to address these issues.

One of the most well-known visual hand gesture recognition methods is color-based gesture recognition because of its computational simplicity and its invariant properties related to the hand shape configurations and due to the human skin-color characteristic values [2]. This method has some drawbacks such as RGB color scheme must be transformed into YUV or HSL space, highly affected by illumination conditions and ineffective in distinguishing objects. To overcome these disadvantages, we propose a hand gesture recognition method using depth data that are robust to light variation and non-intrusive.

Another common approach is to use hand contour extraction [3][4]. In our method, the Moore-Neighborhood algorithm [5] is applied for tracing contour and creating the sorted contour list with high performance. From these contour images, we apply logistic classifier for recognizing hand gesture and obtain high results of both accuracy and real-time performance.
The remainder of this paper is organized according to our novel structure of the major components of proposed hand gesture recognition system. The input image sequence is described in Section 2. Then, section 3 presents proposed hand localization methods. Section 4 explains the feature extraction and classification using hand contour obtained from section 3. Section 5 shows the experimental results of test bed system. Finally, section 6 concludes our proposed method in this paper.

2 Image Sequence Acquisitions

The main purpose of this stage is to produce a sequence of depth motion images that are used for hand localization. First, the depth camera sensor captures approximately 30 depth frames per second and stored in the chronological order. Then the difference image is obtained by subtracting previous image \( I_{t-1} \) from current image \( I_t \), as follows:

\[
\text{Diff}_t = I_t - I_{t-1} \tag{1}
\]

Finally, depth motion image is accumulated from these difference images to represent all movement of human body, hand, object and noise. For removing noise from this motion image we apply median filter and morphological processing [6].

3 Hand Localization Method

Locating hand is one of the most important phases in a hand gesture recognition system. In this section, we set the condition for hand-waving motion which consisting of side-to-side motion image sequence from previous section to initialize hand detection. Then Kalman filter [7] is applied for hand tracking module to determine the trajectory of hand and blob detection [8] is used to separate hand region. Finally, the hand region is extracted from background and exported to binary image that is the input data for gesture recognition stage.

In this step, the modified Moore-Neighbor algorithm is used for hand contour extraction. After clustering the hand position, a group of hand points is found and stored. \( N(a) \) is defined as the eight-point neighborhood of a pixel \( a \); \( p \) represents the current contour pixel and \( q \) represents the starting pixel of current neighborhood checking. \( C \) represents the set of detected contour points, which is initialized to be the empty set. Fig. 1 shows the procedure of this algorithm.

4 Hand Gesture recognition

The human hand is capable of an enormous range of poses. In this research we only focus on 7 poses including 6 poses from no finger to five fingers and OK symbol of sign language.
4.1 Feature Vector Selection

In order to be recognized, the technical feature must model a contour image. Hu invariant moments [9] are chosen to be the feature descriptor. These features are proved invariant to translation, rotation, and scaling. We use Hu invariant moments for the first 7 attributes of feature vector that are invariant to two-dimension transformation. Moreover, we propose the eighth attribute, which presents the relation between the hand region and the contour boundary.

For a hand contour image, the image moment $M_{ij}$ is defined as:

$$M_{ij} = \sum_{x} \sum_{y} x^i y^j I(x,y)$$

The central moment $\mu_{ij}$ is defined as:

$$\mu_{ij} = \sum_{x} \sum_{y} (x-\bar{x})^i (y-\bar{y})^j I(x,y)$$

Where $I(x,y)$ is the intensity at pixel $(x,y)$,

$\begin{align*}
I(x,y) &= \begin{cases} 
1 & \text{if } (x,y) \text{ is on the contour} \\
0 & \text{otherwise}
\end{cases} 
\end{align*}$

i, j are non-negative integers.

---

![Fig. 1 Moore-Neighbor tracing algorithm](image)
\[ \bar{x} = \frac{M_{10}}{M_{00}}, \bar{y} = \frac{M_{01}}{M_{00}} \]

The total area of the objects is given by \( M_{00} \). Scale invariant features can also be found in the scaled central moment. The normalized central moment \( \eta_{ij} \) of order \((i+j)\) is given by:

\[ \eta_{ij} = \frac{\mu_{ij}}{(1+i^2+j^2)M_{00}} \]  

(4)

Based on these definitions, our study computes the Hu invariant moments as follow equations [5 - 12]

\[ H_1 = \eta_{20} + \eta_{02} \]  

(5)

\[ H_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \]  

(6)

\[ H_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \]  

(7)

\[ H_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \]  

(8)

\[ H_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \]  

\[ + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \]  

(9)

\[ H_6 = (\eta_{20} - \eta_{02})[\eta_{30} + \eta_{12} - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12}) \]  

\[ + \eta_{12}(\eta_{21} + \eta_{03}) \]  

(10)

\[ H_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \]  

\[ - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \]  

(11)

\[ H_8 = \sqrt{S/C} \]  

(12)

Where \( S \) represents the square area of the hand, and \( C \) represents the contour’s boundary.

4.2 Training and Classifying

In this research, Logistic classifier [10] used the training data to generate a Logistic regression model with seven labels linked with seven poses. First, each frame’s data is transformed to Logistic regression model data. We define the training data as \( \{y_k, x_i\}_{k=1}^K \); \( y_k \in [1,7]; x_i \in R^m \) is the feature vector that contains 8 attributes as described in previous section. In this case, logistic regression model has K classes (K=7), we basically create K-1 binary logistic regression models where we can choose one class as reference or pivot. Usually, the last class K is selected as the reference. Thus, the probability of the reference class can be calculated by

\[ P(y_i = K|x_i) = 1 - \sum_{k=1}^{K} P(y_i = K|x_i) \]  

(13)

The general form of the probability is

\[ P(y_i = K|x_i) = \frac{\exp(\theta_i^T x_i)}{\sum_{i=1}^{K} \exp(\theta_i^T x_i)} \]  

(14)

As the \( K^{th} \) class is reference \( \theta_K = (0,0,\ldots,0)^T \) and therefore
\[
\sum_{i=1}^{K} \exp(\theta_i^T x_i) = \exp(0) + \sum_{i=1}^{K-1} \exp(\theta_i^T x_i) = 1 + \sum_{i=1}^{K-1} \exp(\theta_i^T x_i) \quad (15)
\]

In the end, we get the following formula for all \( k \leq K \)

\[
P(y_t = k | x_i) = \frac{\exp(\theta_i^T x_i)}{1 + \sum_{i=1}^{K-1} \exp(\theta_i^T x_i)} \quad (16)
\]

This Logistic classifier prediction function is employed to compare this data with the Logistic training model and the result is the pose that is same as the training poses.

5 Experimental results

5.1 Data collection

We captured a hand posture database of a group including three actors for training data. Each hand posture is performed at different positions and angles to Kinect sensor [11]. For each hand posture of a performing actor, about 40 frames of depth images were recorded. Then, the feature vector of each frame were calculated and stored in the hand posture database. Totally, the number of training vectors in traffic gesture database is 840 and each vector is labeled by its posture index number.

5.2 Test-bed system and results

We built a test-bed system by Visual Studio 2010 for evaluating performance of our method. Color image and depth image are retrieved by Microsoft Kinect camera. IKVM library [12] is used to train and test human pose recognition accuracy with Logistic regression classifier. The data set includes 840 samples labeled by seven defined postures. The test mode is 10-fold-cross-validation. Table 1 shows the experimental results (in percentage) for the classifier Logistic regression. The experiments were done on Windows PC of Core i5, and 4GB RAM. The True Positive rate (TP Rate) shows that Logistic Regression classifier returns high result (90.4%). Fig. 2 shows the result of Logistic Regression classifier.

![Fig. 2 Result with Logistic Regression classifier](image-url)
6 Conclusion

In this paper, we proposed a method for human hand gesture recognition detection and tracking based on contour extraction using depth data. In our work, the hand contours are extracted and represented by rotation- and scale- invariant vectors and hand gesture recognition is done by logistic classifier. This method overcomes some drawbacks of most of current hand gesture recognitions. Experimental results from our test-bed system show that it is capable of working in the dark, invariant to user's skin color, clothing and background lighting conditions and it can be easily transplanted to other application.

Acknowledgements. This work was supported by the basic research projects in natural science in 2012 of the National Foundation for Science & Technology Development (Nafosted), Vietnam (102.01-2012.36, Coding and communication of multiview video plus depth for 3D Television Systems)

References