Hand Tracking Method for Multi Modal Interface

Hyun-Koo Kim*         Soo-Young Suk†         Ju H. Park**         Ho-Youl Joug*

* Department of Information and Communication of Engineering, Yeungnam University, Gyeongbuk, Republic of Korea
E-mail: {kim-hk, hoyoul}@ynu.ac.kr

** Department of Electrical Engineering, Yeungnam University, Gyeongbuk, Republic of Korea
E-mail: jessie@ynu.ac.kr

† Gyungbuk Institute of IT Convergence Industry Technology (GITC)
Gyeongbuk, Republic of Korea
E-mail: sysuk@gitc.or.kr

Abstract

This paper proposes hand tracking method using range camera. This camera was located below the hands, facing them. We can detect and track hand to use range-based extraction, connected component analysis process. Our tracking system of hand is more accurate and less sensitive illumination than other hand detection systems. After detection of the hand information from the captured scene, hands are tracked using Kalman filter. Result of hand tracking using our suggested algorithm shows almost 98% accuracy. So, our hand tracking system will be useful in new multi modal interfaces.

Keywords: Hand-detection, hand-tracking, kalman-filter, range camera, multi modal interface.

1. Introduction

Interactions between human and computer are currently performed using keyboards, mice or different haptic devices. In addition to being different from our natural way of interacting, these tools do not provide enough flexibility for a number of applications such as manipulating objects in a virtual environment. In order to improve the human-machine interaction, an automatic hand gesture recognition system could be used. Individual communications are often carried out by means of vocal sounds, gestures, and facial expressions. And body languages also take substantial roles. In many cases, most of information is well included in such actions. Therefore it has been widely studied to extract information from such actions [1, 2]. Relying on visual markings on the hands, previous researchers have recognized sign language and pointing gestures [3, 4]. However, these methods require the placement of markers on the hands. The marking-free systems of [5] can recognize specific finger or pointing events, but not general gestures. Employing special hardware or offline learning, several researchers have developed successful systems to recognize general hand gestures [6]. Blake and Isard [7] have developed a fast contour-based tracker which they applied to hands, but the discrimination of
different hand poses is limited. Extensive research has been conducted on hand gesture recognition making use of digital images [8, 9]. However, it's still ongoing research as most papers do not provide a complete solution to the previously mentioned problems. Range cameras can simultaneously capture a full 3D point cloud with an array sensor at video rates. They offer great potential for real-time measurement of static and dynamic scenes [10]. Investigation on range cameras for automatic gesture recognition is in its infancy. However, some research has been conducted in this area [11, 12]. The final aim of this research is to design and build a human-machine interface using the 3D information provided by a range camera for automatic and real time identification of hand gesture. In this paper, we focus on two applications. The first one is designed to recognize the number of raised fingers that appear in a hand gesture and the second is intended for moving an object in a virtual environment using only a hand gesture on the acquired images. For a real-time application, the expectation is to obtain the best possible images of the hand gesture within the lowest possible time. Some experiments have been conducted with the purpose of defining the best configuration for imaging the hand. This configuration includes, among others, the relative position of the hand and the camera, the influence of the integration time of the camera, the amplitude threshold, and the lighting conditions of the environment, the surrounding objects and the skin colour. This paper is organized as follows: In Section 2, the range camera is presented as well as some of the parameters for defining the best configuration for imaging the user’s hand. Section 3 discusses the multiple steps for hand information extracting and tracking from the image. Section 4 focuses on the obtained results, their analysis, and conclusions.

2. Specifications of Range Camera

The range camera used in this research is the MS 3D depth camera sensor [13]. Once the image is acquired, the range information is used for generating the x, y, z coordinates in meter for each pixel. The range produces images at a rate of up to 30 frames per second.

![Image of range camera](image)

Figure 1. Our multi modal interface system and range camera.

The RGB video stream uses 8-bit VGA resolution (640 × 480 pixels) with a Bayer color filter, while the monochrome depth sensing video stream is in VGA resolution (640 × 480 pixels) with 11-bit depth, which provides 2,048 levels of sensitivity. The range camera sensor has a practical ranging limit of 1.2–3.5 m (3.9–11 ft) distance when used with the Xbox software. Range camera is roughly 6m², although the sensor can maintain tracking through an extended range of approximately 0.7–6 m (2.3–20 ft). The sensor has an angular field of view of 57° horizontally and 43° vertically, while the motorized pivot is capable of tilting the sensor up to 27° either up or
down. The horizontal field of the range camera sensor at the minimum viewing distance of ~0.8 m (2.6 ft) is therefore ~87 cm (34 in), while the vertical is ~63 cm (25 in), resulting in a resolution of just over 1.3 mm (0.051 in) per pixel.

3. Our Proposal Method

In this paper, multiple step based detection has been designed (figure 2). Detection is the process of grouping points that belong to the same object into segments. The idea here is to extract from the point cloud the set of points that describe the hand. The basic idea of detecting the hand area is to segment foreground objects from background objects using their relative distance from the range camera [15].

Foreground objects can be generated by setting the depth measurement window to include the ranges, in which the foreground objects are located. In such a setting, the camera captures light reflected from every object inside the depth measurement window, and ignores light reflected from objects outside the window. Because the 3D camera in our experimental environment set below the hands, facing them, the objects closest to the camera will be hands [14]. Therefore, We assume that the approximate hand area is the predefined hand depth measurement window from the depth of object (hand) closest to the camera. Finally, we obtain a binary image which represents hand parts in a depth measurement window as white pixels, and other components as black pixels. In our experiment, the approximate hand area contains the lower arms as well as the hands. This does not affect with our experiment, because we only want to know the number of hands. After detection of the hand information from the captured scene, hands are tracked by Kalman filter.

3.1. Range-based Extraction

The underlying principle is that there shouldn’t be any object between the camera and the hand. Thus the hand appears in the foreground of the image. Any point between the hand and the camera is considered as a noise. A simple range threshold was used to extract the hand information. The algorithm is designed as follows:

- First step: Find the closest point to the camera using the range.
- Second step: Select points that are less than a threshold from that point.
- Third step: If total number of points lower than a threshold, delete the closest point to the camera and re-start from first step.
3.2. Labeling Method

Connected component labeling is used in computer vision to detect unconnected regions. It is an iterative process that groups neighboring elements into classes based on a distance threshold. A point belongs to a specific class if and only if it is closer within the distance threshold to another point belonging to that same class. After the noise removal, the hand segment appears to be the biggest one in the dataset.

3.3. Hand Tracking

The Kalman filter is a suitable tool designed for this purpose [15]. It is used to predict the position of the hand in the coming frame and after having measured the actual position of the hand in that frame, this prediction is corrected and the adjusted value is used for the prediction in the following frame. The Kalman filter is thus a recursive estimator for linear processes. Position tracking can provide improved results in noisy situations and generate other useful metrics important for the primary system objective. Kalman filtering provides a way to incorporate a linearized version of the system dynamics to generate optimal estimates under the assumption of Gaussian noise. Kalman filtering also provides estimates of state variables that are not directly observable, but may be useful for the system. It is important to have metrics such as rates of change of position robustly estimated. The Kalman filter state variables are updated using the hand-position and angle estimates along with measurements of hand-moving-angle and moving velocity.

\[
\begin{align*}
x_{k+1|k} &= Ax_{k|k} + Bu_k \quad (1) \\
y_k &= Mx_k \quad (2) \\
\end{align*}
\]

Where

\[
x = \begin{bmatrix} \phi, \dot{\phi} = \tan \theta, \ddot{\phi}, W \end{bmatrix}^T \quad (3)
\]

\[
A = \begin{bmatrix} 1 & \nu \Delta t & \frac{(\nu \Delta t)^2}{2} & 0 \\ 0 & 1 & \nu \Delta t & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4)
\]

\[
Bu_k = \begin{bmatrix} 0, \phi \Delta t, 0, 0 \end{bmatrix}^T \quad (5)
\]

\[
M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (6)
\]

These measurements are then used to update the discrete-time Kalman filter for the hand state. The system and measurement equations as well as the Kalman update equations at time \( t \) are detailed in (1)–(6). The initial values for the estimation error covariance and state-noise covariance were determined by empirical testing. The measurement vector consists of the hand position, hand moving angle, and hand size.

4. Test and Conclusion

Identification of hands and distinguishing hands from faces has been slow and difficult in many
algorithms based on color image analysis. We have presented a method for hand detection and tracking based on range data from the camera. Our system detects hand almost 95% when we use not tracking method. But when we use tracking method, our system detects one almost 98%. The utility of the proposed system lies mainly in the enhancement of human-computer interaction in a wide variety of applications. It may be particularly useful in applications in which tactile or verbal interaction is difficult or impossible, including medical operations, industrial and hazardous working environments, and natural interaction with virtual displays.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (20100024741).

References


