Interactive System Based on Hand Tracking : AirGrip

YANG-KEUN AHN, KWANG-SOON CHOI, YOUNG-CHOONG PARK
Korea Electronics Technology Institute
121-835, 8th Floor, #1599, Sangam-Dong, Mapo-Gu, Seoul
REPUBLIC OF KOREA
ykahn@keti.re.kr

Abstract: - This paper proposes a method to control an object on a screen by extracting information about a hand with a depth camera. The proposed method employs an appropriate rectangular region, then extracts the hand by setting the arm as a certain size through vector information. The method employs a convex hull to set a reference coordinate of the hand within the coordinate. The object on the screen is moved or scaled by verifying whether the thumb and index finger are stuck together or not through the extraction of the outline of the hand. The location of the hand is represented through a library capable of implementing 3D. An experiment was conducted with the proposed method and verified the operation of the command on the object.

Key-Words: - Air Grip, Grip, Gesture Recognition, 3D Object Control, Hand Gesture, Space Control

1 Introduction

With their recent development and application in various systems, smart devices are widely employed in diverse locations and uses. The methods used to represent information on display screens are also becoming more diverse. A method commonly used in mobile devices is a touch screen, in which the screen is directly touched to access and control the information. However, such touch screen methods require direct touches and cannot utilize 3D information.

To address these shortcomings, diverse methods have been proposed to utilize a 3D object on a screen with a depth camera.

In conventional research [1], a depth camera is installed on the ceiling facing the floor, and a transparent screen is employed by the hand seen on the other side of the screen. Then the hand and information coordinates are matched on the screen to use the information shown on the screen.

As a alternative, the system proposed by this study installs a depth camera on the floor so that the system can be employed in a place with no ceiling. In addition, a method that differentiates a left hand from a right hand, and a method that employs a convex hull to set a reference point of a hand more effectively than a conventional reference point are also proposed in this paper. An extended Kalman filter was employed to stabilize the coordinate points of the reference point.

The distance between the stabilized reference coordinates and the object is then determined. When the distance between 1 hand and 2 hands gesturing to grab an object and the object is less than a certain distance, the object will move and scale, respectively.

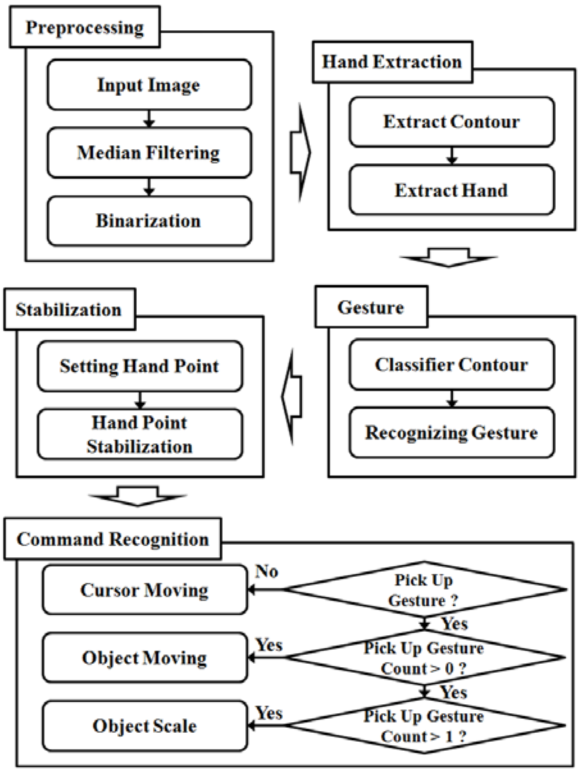


Fig. 1 Flow diagram of the proposed system

2 Pre-process

Because conventional research [1] installs a depth camera on the ceiling, there is a limit in the distance to the floor, but this research installs the camera on the floor. As a result, the usable space becomes infinite, which introduces the possibility that the system might malfunction in an undesirable way.

Such a composition does not ensure the desired control. Therefore, as illustrated in Fig 2, a usable domain of only 30 cm – 60 cm from the camera is set for use.

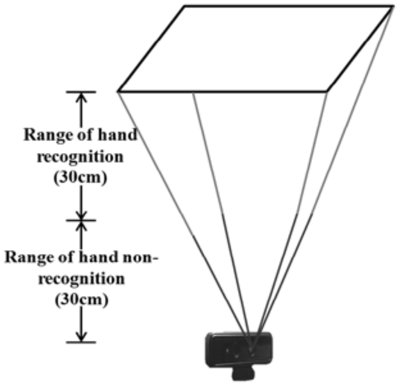

Range of hand recognition (30cm)
Range of hand non-recognition (30cm)

Fig. 2 Usable domain

The data inputted from the camera is the distance data of the object and background as seen from the camera. The inputted distance data is employed as the basis of every image used in the system, and noise reduction is employed to obtain better images. Smoothing computation is conducted for noise reduction. A median filter is employed in the smoothing computation in the proposed system. Data processed out of the usable domain is converted to 0 and not used in the system.

The distance data is converted to have values between 0 - 255 with a proportional expression to primarily produce a black and white image. The implemented black and white image is called a depth image. In the depth image, a light color and high value in a pixel signifies a distance close to the camera, and a dark color and low value in a pixel signifies a distance far from the camera.

An image consisting of the pixels in the usable domain is converted into a binary image. The background in the usable domain is often extracted from an initial image with no hand. In order to remove the background, data of the image from the initial frames with no hand on the screen is accumulated and registered as background values. When the part registered as the background is used, the corresponding part is set to be unusable to prevent its unintended selection.

When the background is removed, undesired parts could also remain in addition to the arm area. An outline is extracted to remove such parts. An open library, OpenCV [3], is employed for the outline extraction [2]. Of the extracted outlines, information of the small outlines are removed from the image.

3 Hand area extraction

Hand and arm areas are extracted from pre-processing. Since the arm area is large, using just the hand area yields a relatively better result. Therefore, the hand area is extracted and an image regarding the hand is created.

First of all, a rectangle most appropriate to the arm is computed. The rectangle is computed with a function provided by OpenCV [3]. A normal vector $\vec{A}_{norm}$ regarding the arm is extracted from the coordinates of the vertices P_1 , P_2 , P_3 , P_4 of the rectangle, and vectors $\vec{A}_{nor\,m_X}$ and $\vec{A}_{nor\,m_Y}$ are generated by projecting the normal vector onto x and y axes.

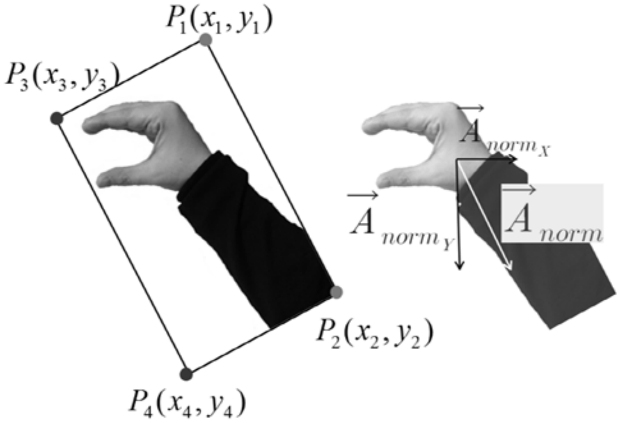


Fig. 3 Rectangle regarding the arm and normal vector

The generated $\vec{A}_{norm}$ is a normal vector that moves along the long side of the rectangle and the directions up and down.

The normal vector has the size of 1 and the sum of $\vec{A}_{nor\,m_X}$ and $\vec{A}_{nor\,m_Y}$ results in $\vec{A}_{norm}$.

$$(1)\ \vec{A}_{norm} = \vec{A}_{nor\,m_X} + \vec{A}_{nor\,m_Y}$$

$$(2)\ \vec{A}_{norm_X} = \frac{(x_2-x_1)}{\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}} = \frac{(x_4-x_3)}{\sqrt{(x_4-x_3)^2+(y_4-y_3)^2}}$$

$$(3)\ \vec{A}_{nor\,m_Y} = \frac{(y_2-y_1)}{\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}} = \frac{(y_4-y_3)}{\sqrt{(x_4-x_3)^2+(y_4-y_3)^2}}$$

Employing the projected vectors $\vec{A}_{nor\,m_X}$ and $\vec{A}_{nor\,m_Y}$ enables the extraction of the area of the hand having a certain size.

P_1 , P_3 and the normal vector $\vec{A}_{norm}$ are employed to determine the values of P_5 and P_6, and the size of the extracted hand area is determined by multiplying α. Therefore, α is the actual height of the extracted hand.

$$(4)\ \ P_5 = P_1 + \alpha \times \vec{A}_{norm}$$

$$(5)\ \ P_6 = P_3 + \alpha \times \vec{A}_{norm}$$

$\vec{A}_{norm}$ is a vector that moves along the sides of the rectangle, thus, P_5 and P_6 are points on arbitrary positions of the sides of the rectangle.

The rectangular area generated by connecting P_2, P_4, P_5 and P_6 is not the hand area and removing this area leads to an approximate extraction of the hand area. On the other hand extraction of a rectangular area by connecting P_1, P_3, P_5 and P_6 leads to the extraction of the hand area.

Setting a range of limitation with the assumption that the left and right arms move in the ranges of 0°~ 90° and 90°~ 180°, respectively, enables $\vec{A}_{nor\,m_x}$ normal vector projected on the x-axis, to be employed in differentiating the left hand from the right hand.

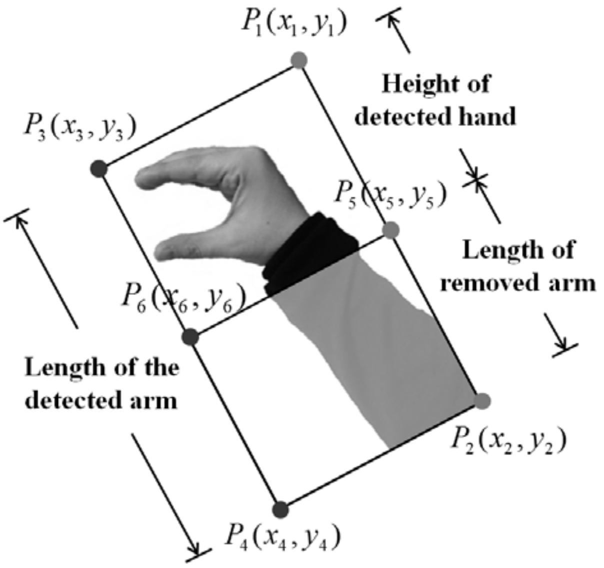


Fig. 4 Extraction of hand area

A hand with positive $\vec{A}_{nor\,m_x}$, left to right direction, is determined to be a right hand, and a hand is determined to be a left hand in the opposite case with right to left direction vector.

$$(6)\ \ Hand = \begin{cases} \text{right} & \left(\text{if}\ \ \vec{A}_{nor\,m_X} > 0\right) \\ \text{left} & (else) \end{cases}$$

4 Gesture classification

The proposed system classifies two types of gestures: a grabbing gesture and not grabbing gesture. As illustrated in Fig 4, the thumb and the index finger of a hand are joined together in a grabbing gesture, and the thumb and index finger are not joined together and simply spread out in a not-grabbing gesture.

The outlines of the hand are extracted from the image to determine a gesture [4]. The extracted outlines should be continuous and outlines of filled and hollow areas should also be extracted.

The extracted areas are checked for included areas, and the areas are categorized into external area and internal area. The outlines are considered to be independent outlines even if an external outline includes internal outlines.

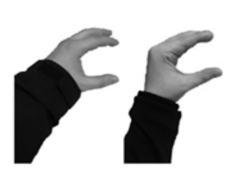
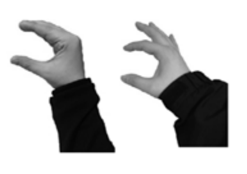
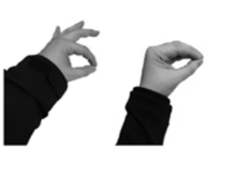
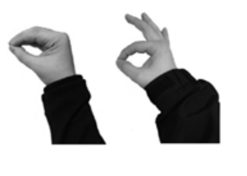
	Left hand	Right hand
Thumb and index finger spread out		
Thumb and index finger joined		

Fig. 5 Gestures used in the system

The hand area with an internal outline signifies a hole created by the joined thumb and index finger of the hand, and the hand area with no internal outline is a hand area with the thumb and index finger spread out. Therefore, an external outline with an internal outline could be considered to be the sign of a gesturing motion.

When there is 1 hand with the grabbing gesture in a certain domain, then an object can be moved. Similarly, the size of the object could be scaled when there are 2 hands with the grabbing gesture.

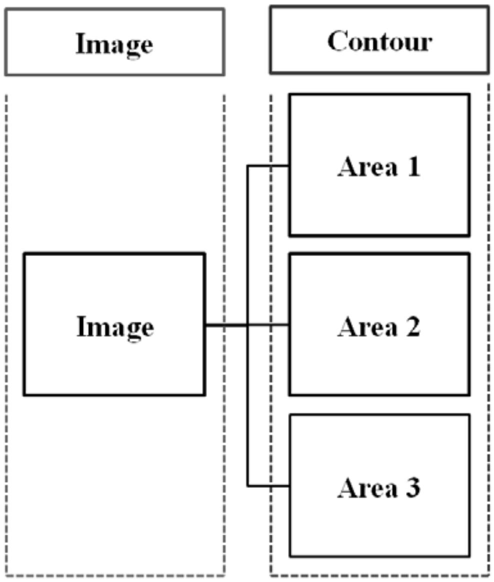


Fig. 6 Outline extraction from the image

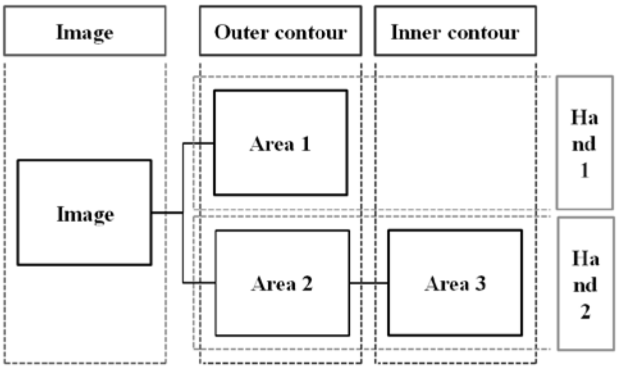


Fig. 7 Categorization of hand areas

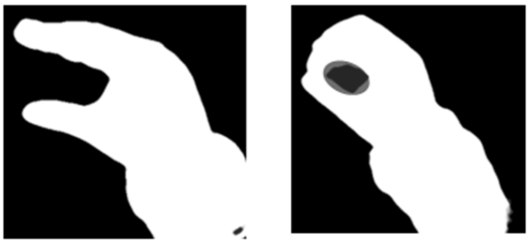


Fig. 8 Determination of the grabbing gesture

5 Determination of the reference coordinate of a hand

The reference coordinate of a hand with an internal outline, grabbing gesture, is different from the coordinate of a hand without an internal outline.

The reference coordinate of a hand with an internal outline, grabbing gesture, is different from the coordinate of a hand without an internal outline.

When the grabbing gesture is not achieved due to the absence of the internal circle, the reference coordinate is determined based on a convex hull and the outline of the hand, and the reference coordinate is generated in between the tips of the thumb and the index finger.

First of all, the outline information of the hand [2] is employed in determining the reference coordinate of the hand, and the outline information is also utilized in the computation of the convex hull [5].

Connecting the points constituting the convex hull reveals that the wrist points are also included in addition to the fingertip points. Consequently, the convex hull is composed of the segments from the outer most points.

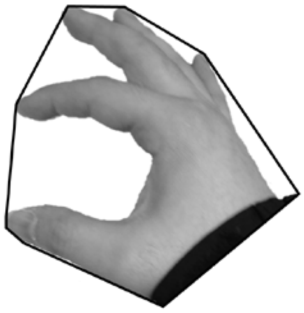


Fig. 9 Convex hull

Computation of a convex defect [6] from the convex hull yields information about the empty space between the convex hull and the image outline. The information of the empty space includes the location of a point on the outline farthest from the convex hull segment and the distance between the segment and the point.

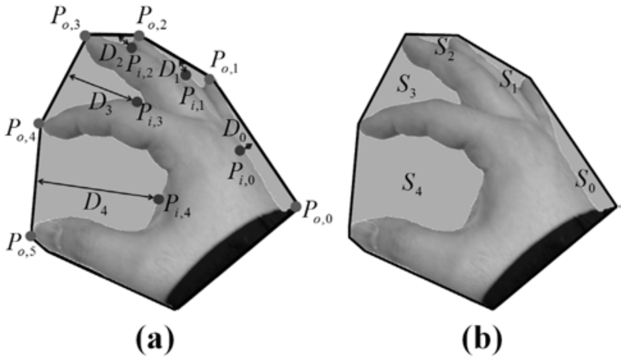


Fig. 10 Information of a hand from convex defect

Connecting the points generated by the convex hull results in the areas illustrated in Fig 10 (a).

P_o points of Fig 10 (b) are the outermost points of the hand generated through the convex hull algorithm. P_i points are the points on the outline farthest from the segments of each convex hull area.

The area S_i with the longest distance between the convex hull segment and the outline is selected, and the center of the segment $P_{o,i}$, $P_{o,i+1}$ corresponding to the area with the longest distance is selected as the reference coordinate of the hand.

The equation of the reference coordinate P_r of a hand with the selected area of S_i is the following.

$$(7)\ P_{r,x} = \frac{P_{o,k,x}+P_{o,k+1,x}}{2}, P_{r,y} = \frac{P_{o,k,y}+P_{o,k+1,y}}{2}$$

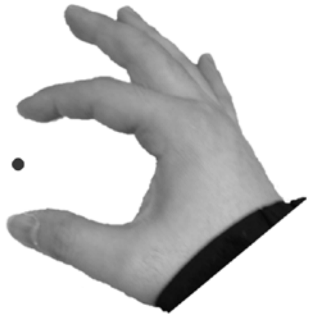


Fig. 11 Reference coordinate of a hand in not-grabbing gesture

The grabbing gesture is a separate difference from the previous case. Similar to the not-grabbing gesture, a convex hull is computed, and the center of the internal area R_c is then computed.

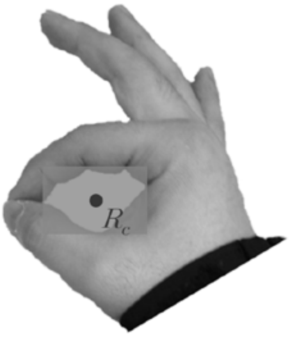


Fig. 12 Center of the internal area in a grabbing gesture

When the convex hull is computed from the grabbing gesture, a fingertip is also extracted from the part where the thumb and the index finger overlap, and the selection of this point is an essential issue. The Euclidean distances between the extracted fingertips $P_{o,j}$ and the center of the internal area $R_c$are measured for the selection of the point.

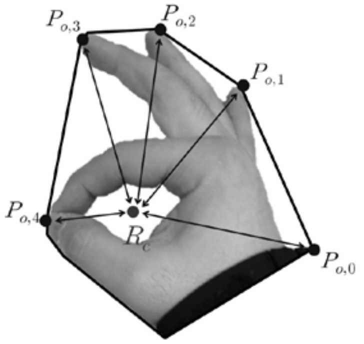


Fig. 13 Distance measurement between the extracted fingertips and the center of the internal area

$$(8)\ D_j = \sqrt{\left(P_{o,j,x} - R_{c,x}\right)^2 + \left(P_{o,j,y} - R_{c,y}\right)^2}$$

Choosing the fingertip with the shortest measured distance results in the selection of the joint part of the thumb and the index finger.

Coordinates from the image coordinate system are used for the x, y coordinates of the selected reference coordinates, and the unit of the coordinates is a pixel.

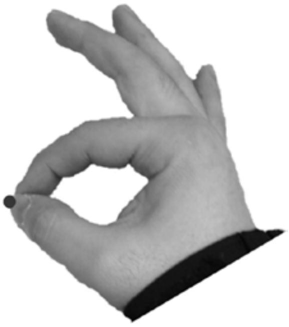


Fig. 14 Reference coordinate in a grabbing gesture

The z coordinate of the reference may or may not be obtained. Therefore, the closest z coordinate of the hand is used and the unit of the z coordinate is mm. The location of the hand is represented with the obtained coordinates.

6 Domain of utilization of the hand on the screen

When a hand moves out of the upper view angle of the camera, the hand is not continuously captured by the camera, but the arm is continuously captured by the camera, and the system cannot properly recognize a gesture. The domain of the utilization of the hand is slightly decreased to prevent such a malfunction.

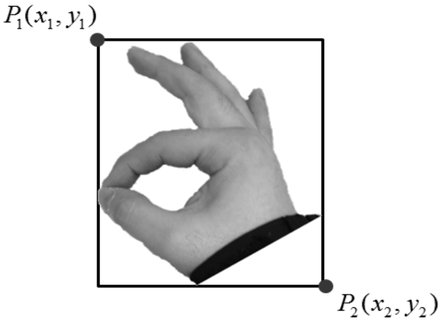


Fig. 15 Rectangle for the outer area of a hand

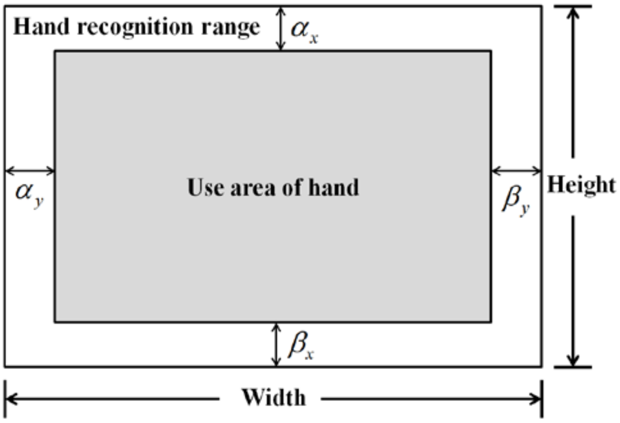


Fig. 16 Domain of utilization of a hand

A rectangle is constituted for the outer area of the hand, and the hand is utilized only when the conditions within the domain of the utilization are satisfied. On the other hand, information of the hand is removed when the conditions are not satisfied, to prevent undesired computation

The conditions within the domain of utilization are computed with the two outermost points of the outer rectangle, and the conditions could also be computed with the other two points of the rectangle.

Conditions within the domain of utilization :

$$(9)\quad \boxed{\begin{array}{c} \alpha_x < x_1 \\ x_2 < Width - \beta_x \\ \alpha_y < y_1 \\ y_2 < Height - \beta_y \end{array}}$$

6.1 Stabilization of the coordinates of a hand

The x and y coordinates of the coordinate system of the image are employed as the coordinates of a hand, and the distance measured by the depth camera is employed as the z coordinate. However, the values have large noise. Therefore, an extended Kalman filter [7] is used to stabilize the coordinates. A state model and a measurement model are required for the use of an extended Kalman filter.

State model :

$$(10)\ x_t = f(x_{t-1}) = \begin{bmatrix} x_t \\ y_t \\ z_t \\ \theta_t \\ \varphi_t \\ d_t \end{bmatrix} = \begin{bmatrix} x_{t-1} + d_{t-1} \times \sin\theta_{t-1} \times \cos\varphi_{t-1} \\ y_{t-1} + d_{t-1} \times (-\sin\theta_{t-1}) \\ z_{t-1} + d_{t-1} \times \cos\theta_{t-1} \times \cos\varphi_{t-1} \\ \theta_{t-1} \\ \varphi_{t-1} \\ d_{t-1} \end{bmatrix}$$

Measurement model :

$$(11)\ z_t = h(x_t) = \begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \\ z_t \\ \theta_t \\ \varphi_t \\ d_t \end{bmatrix}$$

x, y, and z represent the coordinates of the hand; d represents the distance between the hand and the origin; φ represents the angle regarding y and z of the hand on Y-Z plane; θ represents the angle regarding x and z on the X-Z plane.

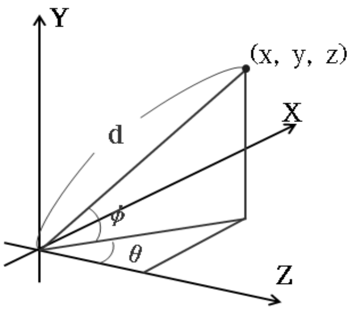


Fig. 17 Model for the employment of extended Kalman filter

7 System composition

3D objects and a program that expresses the location of the hands were implemented to verify the operation of the system proposed by this paper. In the system, the gesture of a user's hand is inputted into a PC through a depth camera. Then, the PC analyzes the gesture and moves or scales the object shown on the screen.

7.1 Hardware composition

Fig 18 illustrates the composition of the hardware. The hardware is composed of a depth camera, a transparent monitor, and a PC. The depth camera captures the hand and transmits the image to the PC through a USB interface. The PC processes and treats the image recorded by the depth camera for the gesture analysis. The information processed through the gesture analysis is shown in the transparent monitor through HDMI interface.

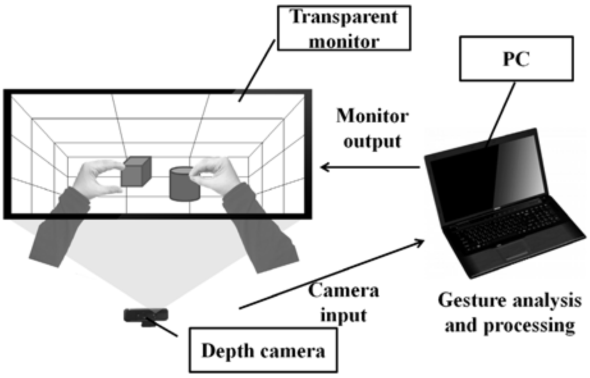


Fig. 18 System hardware composition

7.2 Software composition

The software proposed by this study is largely categorized into two parts. The first part is the process part that processes a given gesture, and the second part is the expression part that expresses the control of the objects with the gestures. The process part shows the operation of the proposed algorithm, and the expression part shows the motion and size scaling of the object according to the location and gesture of the hand.

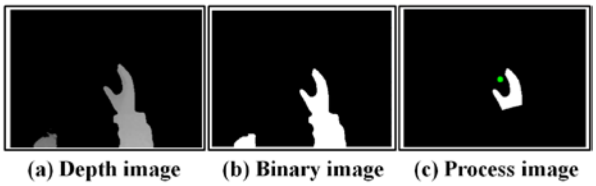


Fig. 19 Images of the process part

8 Experiment results and conclusion

The experiment was conducted in an environment where a depth camera, DS325, was connected to a PC with Intel (R) Core (TM) i7-2600K 3.4GHz CPU and 8Gbyte memory

The program presented a 3D object with respect to each plane, the X-Y plane, Y-Z plane, and X-Z plane, executed appropriate commands according to the gestures, and expressed the location of the hand to acknowledge the location of the hand in the program.

The coordinate system that was actually measured is the coordinate system of the left hand, and the expression program employing the right hand coordinate system did not yield a satisfactory result. Therefore, the values of the y axis were reflected by multiplying negative values to convert the left hand coordinate system to the right hand coordinate system. In addition, the actually measured coordinate system is the coordinate system of the camera with the camera as the origin.

The command to the object is categorized by the number of hands exhibiting a grabbing gesture within a certain distance from the object. The object can be picked up and freely moved around when there is 1 hand with the grabbing gesture within a certain distance from the object. When there are 2 hands with the grabbing gesture within a certain distance from an object, the object's size would be scaled. The object will be smaller when the distance between the two hands decreases, and be larger when the distance between the two hands increases.

The operation of the gesture was verified through the program.

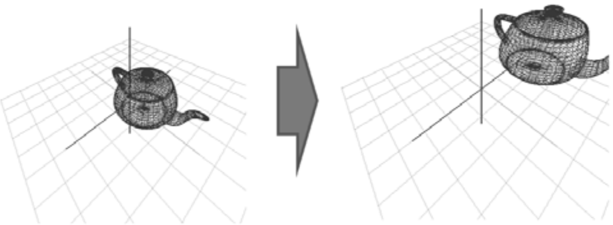


Fig. 20 Movement of an object

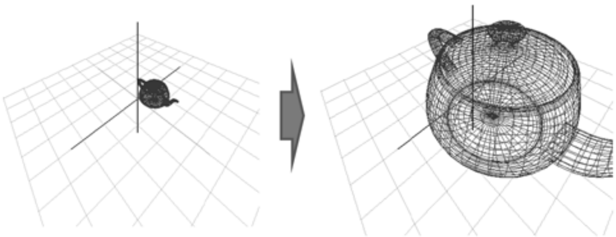


Fig. 21 Scaling of an object

References:

[1] Jinha Lee, Alex Olwal, Hiroshi Ishii, and CatiBoulanger, “SpaceTop :Intergration 2D and Spatial 3D Interaction in See-through Desktop Environment,” In Proceeding of SIGCHI, 2013, pp. 189-192.

[2] Suzuki. S. and Abe. K, "Topological Structural Analysis of Digitized Binary Images by Border Following," CVGIP, 1985, pp. 32-46.

[3] Intel Corporation. Open Source Computer Vision Library reference manual, Dec. 2000.

[4] Andrew D, "Using a depth camera as a touch sensor," ACM International Conference on Interactive Tabletops and Surfaces, 2010, pp. 69-72.

[5] Sklansky. J, "Finding the convex hull of a simple polygon," PRL 1 $number, 1982, pp. 79-83.

[6] K. Homma and E. I. Takenake, "An image processing method for feature extraction of space-occupying lesions," Journal of Nuclear Medicine 26, 1985, pp. 1472-1477.

[7] Grewal, Mohinder, "Kalman filtering : Theroy and Practice Using Matlab," John Wiley & Sons Inc, 2008.