3D-Position Estimation for Hand Gesture Interface Using a Single Camera

Seung-Hwan Choi, Ji-Hyeong Han, and Jong-Hwan Kim

Department of Electrical Engineering, KAIST,
Gusung-Dong, Yusung-Gu, Daejeon, Republic of Korea
{shchoi,jhhan,johkim}@rit.kaist.ac.kr

Abstract. The hand gesture interface is the state of the art technology to provide the better human-computer interaction. This paper proposes two methods to estimate the 3D-position of the hand for hand gesture interface using a single camera. By using the methods in the office environment, it shows that the camera is not restricted to a fixed position in front of the user and can be placed at any position facing the user. Also, the reliability and usefulness of the proposed methods are demonstrated by applying them to the mouse gesture recognition software system.

Keywords: Position Estimation, Hand Gesture Interface, Human Computer Interaction.

1 Introduction

Recently, various hand gesture interfaces have been developed. Among them, the vision-based hand gesture interface with a webcam is efficient and popular because it only needs a webcam which is already very common. There are mainly two approaches in the vision-based hand gesture interface, i.e. 3D-model-based and 2D-appearance-based. 3D-model-based approach fits the hand image to a 3D hand model, which is already constructed, using extracting hand features or hand outlines from the camera image [1][2][3]. This approach estimates the hand pose highly accurately, but it takes a long processing time. 2D-appearance-based approach directly compares the input hand image to the database images [4][5]. This approach takes a short processing time, but it does not estimate the hand pose accurately when the wrist or the forearm is moving. The hand pose is easily coupled with information such as software commands like click or drawing. However, users have to study each hand pose corresponding to each command to use the software.

The hand position, on the other hand, is more intuitive to control the software than the hand pose. Additionally, most of software is also available to use the interface with the 3D hand position because the 3D-position can be easily translated to a 2D-position and additional state information on mouse commands. There are several methods to find out the 3D-position, such as by using a depth cam or a stereo cam, but they are still not popular among most of the users. This paper proposes two different methods to detect the 3D-position of a hand by using a single camera for the hand

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2011, LNCS 6762, pp. 231–237, 2011.

gesture interface. The proposed methods use homography and the neural network. The camera is not restricted to a fixed position in front of the user and can be placed at any position facing the user.

This paper is organized as follows: Section 2 describes the hand detection and proposed methods for estimating 3D-position of the detected hand. In Sections 3, the experimental results are described. Finally, concluding remarks follow in Section 4.

2 3D-Position Estimation

The proposed method to estimate 3D-position of the hand consists of two processes. The first one is hand detection in a camera image. The camera image contains information about the hand such as its 2D-position, direction and pose. Among them, the 2D-position and the size of the hand are used. The other one is the 3D-position estimation of the detected hand. This paper proposes two methods which are based on the homography and the neural network, respectively, for estimating the 3D-position of the hand from a camera image.

2.1 Hands Detection

The color based object recognition is used to detect the hand in a camera image. A red colored glove is used in this research to help distinguishing the hand from the other objects in environment. Fig. 1 shows the binary image with the number of pixels of detected hand. The size of an object in the image has the information about the distance on z-axis if it is spherical like a ball. Even though the hand is not spherical, the size of the hand is still useful information to find the distance on z-axis in a 2D-image.

Fig. 1. The result of detected hand

2.2 Homography-Based Method

Homography has been used in the field of computer vision. Any two images of the same planar surface in space are related by homography. If there is proper initial processing to find out the relation between the distorted 3D-position in the camera image and the 3D-position in the real world, the homography matrix can be calculated. The eight positions are sampled for initialization, which is described in the following.

1. The user imagines a hexahedral available space for the hand gesture interface.
2. The user moves the hand to each corner of the available space and inputs the corresponding position information one by one, as shown in Fig. 2.
3. Calculating the homography matrix between the image space and the real world by using singular value decomposition:

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = H \begin{bmatrix} u \\ v \\ w \\ 1 \end{bmatrix} \tag{1}$$

where H is the homography matrix, (u,v) is the position in the image space, w is the size of the hand in the image space, and (x,y,z) is the position in the real world.

Fig. 3 shows eight input positions and sizes of the hand in the image space according to each marker. The radius of each circle is the same as the square root of the number of pixels of detected hand.

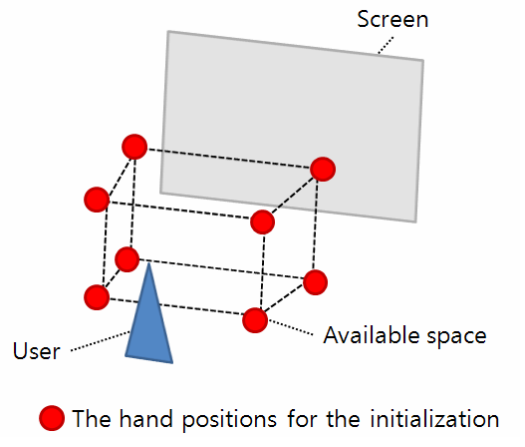

Fig. 2. Initialization for homography method

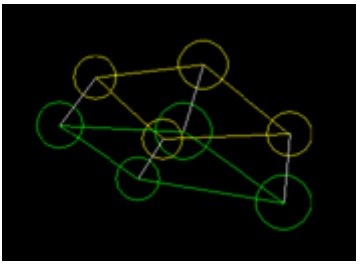

Fig. 3. The input data for homography method

2.3 Neural Network-Based Method

The size of the hand in the camera image does not guarantee the depth information because the hand is not spherical. The pose and the size of the hand in the camera image are almost the same when it returns to the same position while a user moves it freely. The neural network can be trained by the relation between the distorted 3D-position in the camera image and the 3D-position in real world. The initialization process is described in the following.

1. The user imagines a hexahedral available space for the hand gesture interface.
2. The initialization program shows a moving marker on the screen as shown in Fig. 4. This is the target data of the neural network. The user moves the hand continuously to proper positions according to the marker. This is the input data of the neural network (It takes about 15~25sec.),
3. Training the neural network (It takes about 10~15sec.).

Fig. 5 shows the input positions and sizes of the hand in the image space according to the moving marker.

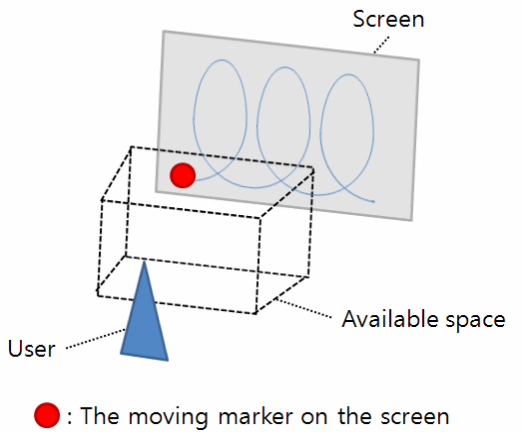

Fig. 4. Initialization for the neural network method

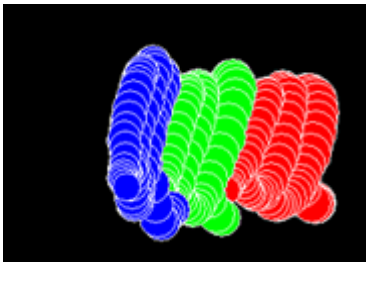

Fig. 5. The training data for the neural network method

3 Experimental Results

3.1 Position Error in MS Windows Environment

Both methods were tested four times and the camera was relocated at each time as shown in Fig. 6. The test program drew 50 target positions on the screen one by one and each target position was generated randomly. It showed each target position during one second and users pointed the target position in their own interface space. It saved the estimated 3D-position of the hand when it changed the target position to the next one. The error is calculated by the distance between the target position and the estimated 3D-position of the hand in the x-y plane only, since it was tested in the MS windows environment. The position space was normalized [-1,1] on each axis. Table 1 shows the average errors and the standard deviations of the experiments.

Table 1. The experimental results

Experiment #		1	2	3	4
Homography	AVG	0.1106	0.0764	0.0797	0.1864
	STD	0.4317	0.1157	0.1948	0.2897
	%	**21.59**	**5.78**	**9.74**	**14.48**
Neural network	AVG	0.0917	0.0839	0.0862	0.0888
	STD	0.2318	0.2947	0.1504	0.1769
	%	**11.59**	**14.73**	**7.52**	**8.84**

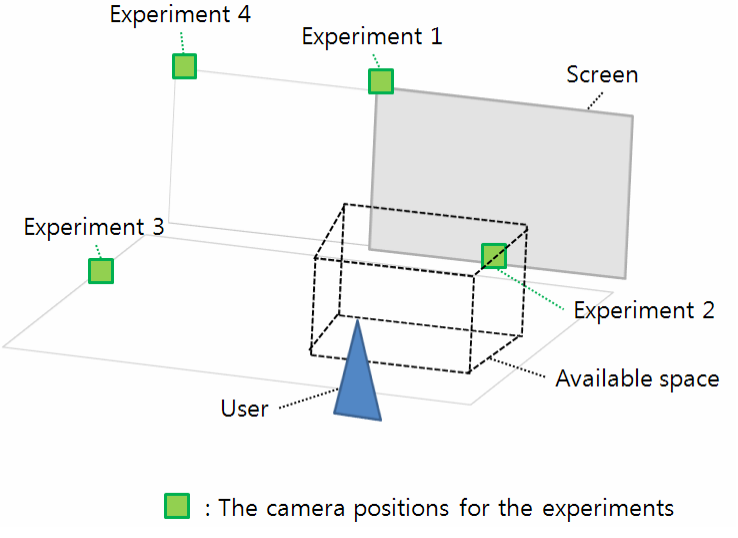

Fig. 6. The experimental environment

The method using the neural network generally showed better performance than the one using homography. The one using homography was sensitive to the location of the camera because the hand is not spherical. In summary, the method using the

neural network is better to estimate the 3D-position of the hand by using a single camera. Exceptionally, the method using homography is more efficient than the one using the neural network when the camera is located in front of the user or a spherical object is used instead of a hand.

3.2 Applying to Mouse Gesture Recognition Software

The proposed methods were tested with the Stroke-It; a mouse gesture recognition software [6]. If the position value of the hand on the z-axis was lower than the predefined threshold, it was interpreted as a pressed down mouse button (Fig. 7). The user held down the mouse button by a hand gesture and then drew the gesture. Once the gesture was performed, the software would execute the action associated with the gesture. The mouse gesture recognition software executed the web browser when the mouse drew 'w' (Fig. 8).

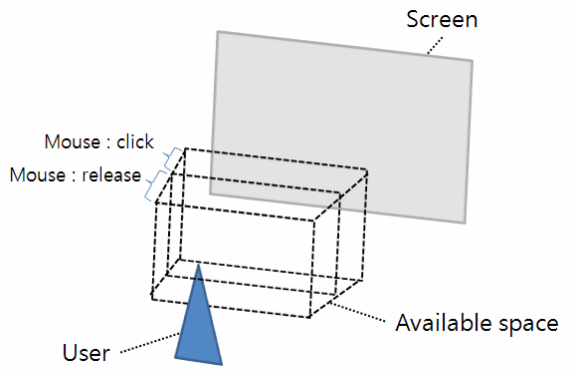

Fig. 7. The experiment to control the cursor of the mouse

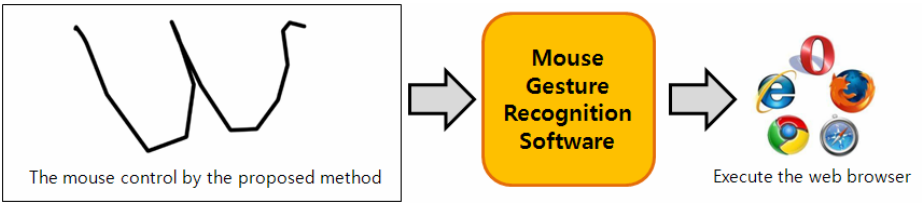

Fig. 8. The experiment with the mouse gesture recognition software

4 Conclusion

This paper dealt with the 3D-position estimation of the hand using a single camera. The two methods using homography and the neural network were proposed to estimate the 3D-position of the hand from a camera image. They were both implemented and tested with the camera at not only in front of the user but also

various locations. The experiment with the mouse gesture recognition software showed that the proposed method could be easily utilized in the office environment or in the presentation room. As the future work, the other information of the hand in a camera image should be applied to the neural network method to improve the performance.

Acknowledgments. This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the National Robotics Research Center for Robot Intelligence Technology support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-N02100128).

References

1. Hoshino, K., Tomida, M.: 3D Hand Pose Estimation Using a Single Camera for Unspecified Users. Journal of Robotics and Mechatronics 21(6), 749–757 (2009)
2. Ueda, E., Matsumoto, Y., Imai, M., Ogasawara, T.: Hand Pose Estimation for Vision-based Human Interface. IEEE Transactions on Industrial Electronics 50(4), 676–684 (2003)
3. Jeong, M.H., Kuno, Y., Shimada, N., Shirai, Y.: Recognition of Two-Hand Gesture Using Coupled Switching Linear Model. IEICE Transactions on Information and Systems E86-D(8), 1416–1425 (2003)
4. Athitos, V., Scarloff, S.: An Appearance-based Framework for 3D Hand Shape Classification and Camera Viewpoint Estimation. In: Proc. IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, pp. 40–45 (2002)
5. Wu, Y., Lin, J., Huang, T.S.: Analyzing and Capturing Articulated Hand Motion in Image Sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(12), 1910–1922 (2005)
6. StrokeIt, `http://www.tcbmi.com/strokeit/`