J-013

# Gesture Recognition with Both Hands and Head Pose by using Depth Sensor

Graciliano Garcia Torres Galindo Jr † Chendra Hadi Suryanto† Kazuhiro Fukui†

**Abstract**

This paper proposes both hands and head pose recognition for human-machine interaction system. The goal of our system is to provide more natural interaction than conventional methods by considering the head pose. In the system, we used the set of sequential depth images obtained by using SoftKinetic DepthSense 325 for the hand shape recognition. On the other hand, grayscale images are used for head pose recognition. In this initial work, the recognitions for hands and head pose are done separately, by adopting the kernel orthogonal mutual subspace method framework.

## 1 Introduction

Due to the advancement in robotics and computer visions, natural interaction between human-machine is highly demanded for various applications. For example, to operate robots underwater, hard terrain, or other dangerous or contaminated area. Most of the vision based conventional systems teleoperate robot by just using the hand gesture information. For example, in [1], four gestures of hands obtained by using Kinect were used to command robots. In [2], both hand gestures are captured with Kinect. Since both-hands are used, more commands can be incorporated, resulting 18 commands for controlling the robots.

In this paper, we propose a system that incorporates not only both hands, but also the head pose to induce commands. The commands consist of pointing directions with the right hand and setting the amount of the movements with the left hand. The head pose is used as additional condition whether the command should be executed or not. We adopted the framework of the kernel orthogonal mutual subspace method [4] for the classification, the high classification performance of which has been confirmed [2, 3, 5].

The organization of this paper is as follows. First we describe the basic idea in Section 2. Then, in Section 3 we describe the framework of our system. In Section 4, the results of initial experiments are reported. Finally, Section 5 provides conclusions and our future works.

## 2 Basic Idea

The main objective of our work is to introduce a robot control system with more natural interaction. The basic idea is that the user will utilize the left hand to use one of five commands, which are to show one, two, three, four or five fingers to indicate the magnitude of the intended movement. As for the right hand, the commands are to point to one of the four directions (move forward, turn left, move backward, turn right) or to signalize the stop action. But these actions are not intended to be done every time the user issues the commands. These commands will only be applied depending on whether the user is gazing at the camera or not. In this sense, we also detect the head pose to recognize if it is at frontal position or non frontal position. In practice, this can be used in a situation where there are lots of instances of the system, and one will only work when the action is intended to it. By the combination of the 5 gestures from each hand, our system can deal up to 25 gestures with additional state from the head pose.

In this early work, the recognition of each hand and the head pose were done separately. The use of sequential depth image can output a high performance hand shape recognition system, such as in [3]. This motivates us to adopt the same method with [3] for the hand shape recognition. On the other hand, for head pose recognition, we simply used the sequential grayscale face images. After the set of depth images from left hand, right hand, and grayscale of face images are collected, nonlinear orthogonal subspaces are generated for each of them. Then the recognition is performed by computing the similarity between the corresponding training subspaces and the test subspaces. Finally, the system selects a corresponding command to the gesture class which has the highest similarity.

## 3 Proposed Framework

Figure 1 shows the overview of our framework. Basically, our framework adopts similar approach to [3, 2] with the additional head pose recognition.

### 3.1 Training phase

Since our system combines all the three types of gestures, i.e., left hand, right hand, head pose, it is possible to conduct the training phase separately for each type. The process flow

†University of Tsukuba, Graduate School of Systems and Information Engineering, Department of Computer Science

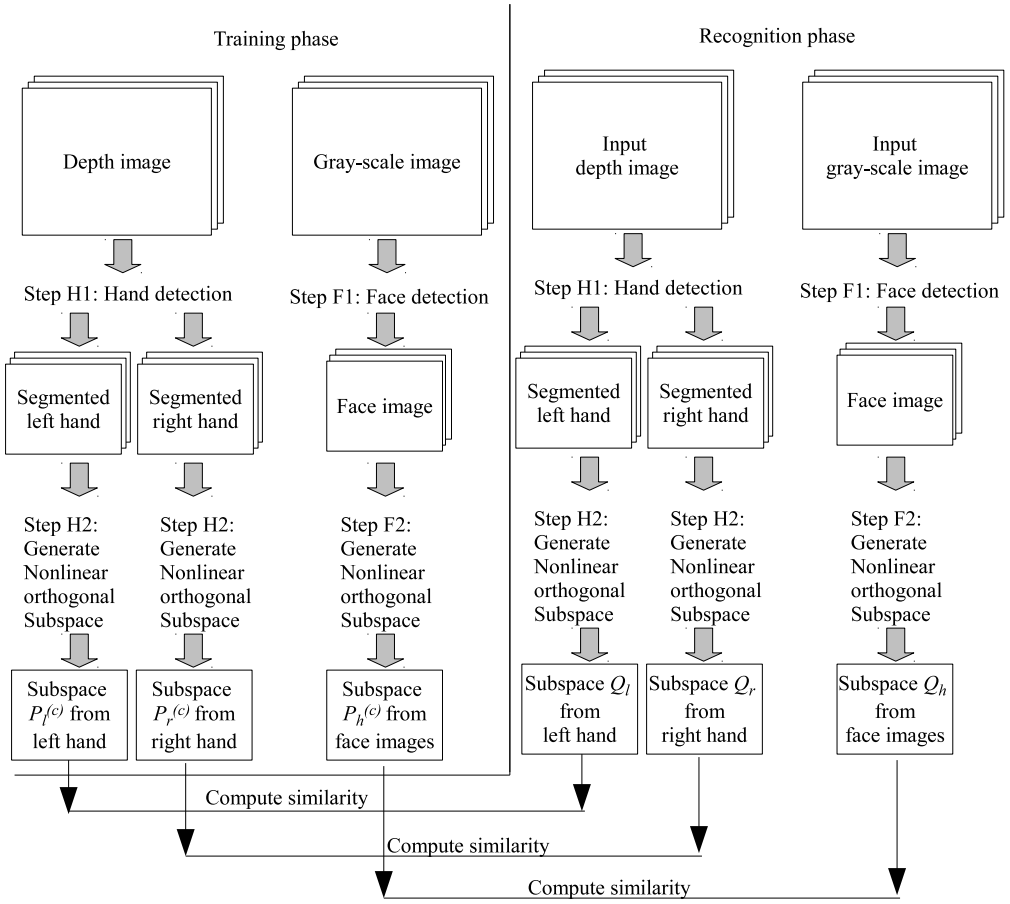


Figure 1: General overview of the framework.

of the training phase for left and right hands is as follows:

**Step H1:** Given a depth image obtained by using DepthSense 325, apply hand detection. To simplify the process, two assumptions are made: the hands are the nearest objects to the depth sensor camera; and the left hand is positioned at the left region of the given depth image, while the right hand is positioned at the right region of the given depth image.

**Step H2:** By using the framework of kernel orthogonal mutual subspace method (KOMSM) [4], nonlinear orthogonal dictionary subspace is generated for each set of the vectorized depth images from each class. Since there are 5 classes of gestures for each hand, there are 10 dictionary subspaces generated.

The process flow of the training phase for head pose is as follows:

**Step F1:** Given a grayscale image obtained by using DepthSense 325, apply face tracking. We simply used the face tracking function provided by OpenCV [6].

**Step F2:** Similar to the hand shape, a dictionary subspace for each head pose is generated by applying KOMSM framework to the set of the vectorized gray scale images. As we only have two classes (frontal and non-frontal), there are two dictionary subspaces for head pose.

From the training phase, we obtained the set of dictionary subspaces $\{\mathcal{P}_l^{(1)}, \ldots, \mathcal{P}_l^{(5)}\}$ for the left hand gestures, $\{\mathcal{P}_r^{(1)}, \ldots, \mathcal{P}_r^{(5)}\}$ for the right hand gestures, and $\{\mathcal{P}_h^{(1)}, \mathcal{P}_h^{(2)}\}$ for the head pose of frontal and non-frontal.

## 3.2 Recognition phase

In the recognition phase, the step for generating the input subspaces are similar to the training phase. First, input subspaces $\mathcal{Q}_l$, $\mathcal{Q}_r$, $\mathcal{Q}_h$ are generated from left hand, right hand, and head pose, respectively. Then, the similarity between each pair of left hand gesture $Sim(\mathcal{P}_l^{(c)}, \mathcal{Q}_l)$, right hand gesture $Sim(\mathcal{P}_r^{(c)}, \mathcal{Q}_r)$, and head pose $Sim(\mathcal{P}_h^{(c)}, \mathcal{Q}_h)$ are computed, where $c$ here indicates the class for the left hand, right hand, and head pose. $Sim(\mathcal{P}, \mathcal{Q})$ denotes the function that computes the similarity between two subspaces. Finally, the set of the input gestures are classified into the class which has the highest similarity.

Non frontal face

Frontal face

Figure 2: Samples of frontal and non-frontal face images.

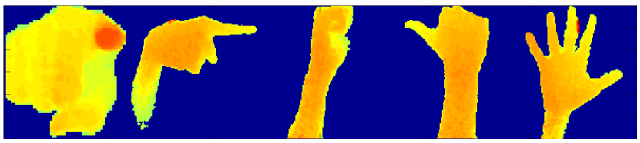

Figure 3: Five classes of right hand gestures. From left to right: forward, left, backward, right, and stop.

## 4 Evaluation Experiments

In this early work, the recognition experiments were conducted offline by using Matlab*. The dataset for the experiment were collected as follows. First, 30 frames of images for each class were collected 10 times. Then, each image is resized to 30 × 30 pixels. Figure 2 shows some examples of the collected face images. The example of the collected depth images from right hand and left hand for each class are shown in Figures 4 and 3, respectively.

The parameters for KOMSM were set as follows. The reference subspace dimension was fixed to 15, while the input subspace was fixed to 4. The Gaussian bandwidth kernel parameter for the KOMSM was set to 0.1. The evaluation was done by using leave-one-out scheme, where one set is used for testing and the rest were used for training. Table 1 shows the recognition rate for each type of gestures when using 5, 10, 15, and 30 sequential images. This results suggest that by using more images we can obtain better performance. We obtained 100% recognition rate for the head pose when using 30 sequential images. However, we could only obtain 98% recognition rate for the left hand gestures. While for the right hand gestures, we could easily achieve 100% recognition rate with few number of images.

To further understand which gesture from the left hand is difficult to recognize, Tables 2 and 3 display the confusion matrices of the left hand gesture recognition when using 10 and 30 sequential images, respectively. From both Tables, we can see that classes 2 and 3 were sometimes mistakenly recognized to each other. While class 5 was also mistakenly classified to class 4. One possible reason of the misclassification is because of the imperfect segmentation of the hand shape, since we used a very simple approach to automatically crop the hand. Besides, we used a vectorized data of the depth data, which means that it can be sensitive to the variation of the hand pose and position.

*We used mexopencv for the face tracking on the Matlab. URL: http://www.cs.stonybrook.edu/~kyamagu/mexopencv/

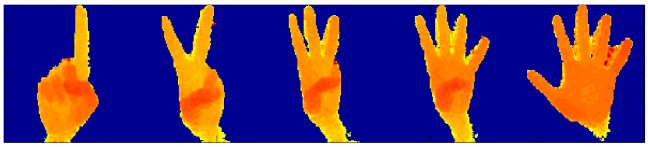

Figure 4: Five classes of left hand gestures to indicate the magnitude of the movements based on the right hand command.

| Gesture types/Num. of frames | 5 | 10 | 15 | 30 |
|---|---|---|---|---|
| Head pose | 85 | 90 | 95 | 100 |
| Left hand | 90 | 90 | 98 | 98 |
| Right hand | 100 | 100 | 100 | 100 |

Table 1: Recognition rate for each type of gestures (in %) using 5, 10, 15, and 30 frames of sequential images.

| | class 1 | class 2 | class 3 | class 4 | class 5 |
|---|---|---|---|---|---|
| class 1 | 10 | 0 | 0 | 0 | 0 |
| class 2 | 0 | 9 | 1 | 0 | 0 |
| class 3 | 0 | 2 | 7 | 1 | 0 |
| class 4 | 0 | 0 | 0 | 10 | 0 |
| class 5 | 0 | 0 | 0 | 1 | 9 |

Table 2: Confusion matrix for the left hand gestures when using 10 sequential images.

| | class 1 | class 2 | class 3 | class 4 | class 5 |
|---|---|---|---|---|---|
| class 1 | 10 | 0 | 0 | 0 | 0 |
| class 2 | 0 | 10 | 0 | 0 | 0 |
| class 3 | 0 | 0 | 10 | 0 | 0 |
| class 4 | 0 | 0 | 0 | 10 | 0 |
| class 5 | 0 | 0 | 0 | 1 | 9 |

Table 3: Confusion matrix for the left hand gestures when using 30 sequential images.

## 5 Conclusion

We proposed a framework that provides a more natural interaction with robots or machines, which can be operated using both hands while considering the head pose as well. The initial experiments showed a promising results, where the recognition rate achieved up to 98% for the left hand and 100% for the head pose and the right hand.

In the future, we will integrate the three types of the gestures (left hand, right hand, and head pose) into a real-

time recognition system. Moreover, to further improve the recognition performance, we will consider to adopt feature extraction methods for the depth and gray scale images, especially to deal with the variation of the position and scale of the hands.

## Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 24300290.

## References

[1] Kun Qian, Jie Niu, and Hon Yang “Developing a Gesture Based Remote Human-Robot Interaction System Using Kinect”, International Journal of Smart Home, Vol.7, No.4, pp.203–208, 2013.

[2] Martin Peris Martorell and Kazuhiro Fukui, “Both-hand Gesture Recognition Based on KOMSM with Volume Subspaces for Robot Teleoperation”, IEEE-Cyber, pp.192–196, 2012.

[3] Daisuke Takabayashi, Yoto Tanaka, Akio Okazaki, Nobuko Kato, Hideitsu Hino, and Kazuhiro Fukui, “Finger alphabets recognition with multi-depth images for developing their learning system”, 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV), 2014.

[4] Kazuhiro Fukui and Osamu Yamaguchi, “The Kernel Orthogonal Mutual Subspace Method and its Application to 3D Object Recognition”, 8th Asian Conference on Computer Vision (ACCV), LNCS, Vol.4844, pp.467–476, 2007.

[5] Yasuhiro Ohkawa and Kazuhiro Fukui, “Hand Shape Recognition Using the Distributions of Multi-Viewpoint Image Sets”, IEICE Transactions on Information and Systems, 95-D(6), pp.1619–1627, 2012.

[6] Gary Rost Bradski “The OpenCV Library”, Dr.Dobb’s Journal of Software Tools, 2000.