Segmentation of Continuous Human Gestures for Imitative Learning

Hyoungnyoun Kim Kyungwha Park Kyungkoo Kim Ji-Hyung Park
Intelligence and Interaction Research Center, Korea Institute of Science and Technology
{nyoun; kh_park; kkkim; jhpark}@kist.re.kr

Abstract − Imitative learning, which teaches human gestures to robots by demonstration, is an active field of research in humanoid robotics. In this paper, in order to allow robots to learn complex/continuous gestures, we propose a method for segmenting and recognizing such gestures based on predefined basic motions. Since each elemental gesture has ambiguous data ranges in a sequential gesture, we combined both Dynamic Time Warping and Kullback-Leibler divergence to increase the accuracy of gesture segmentation. We have achieved about 80 percent accuracy in segmenting continuous gestures, and applied our approach to a small humanoid robot to demonstrate that imitative learning is possible.

Keywords − Gesture recognition, imitative learning, Dynamic Time Warping.

1. Introduction

Recently, many researchers have been interested in imitation learning as one of the methods for learning human gestures in human-robot interaction. Imitative learning of robots has been accomplished by tracking the body using vision capture devices or by corresponding the joint angles using motion sensors [1,2,3]. However, the gestures that a robot performs are limited to those that have been pre-programmed. Since most robot applications have constrained and predefined mobility, learning new gestures is important to increase the capability of a robot. Furthermore, if a robot can learn undefined gestures by observation, it will have greater applicability in human-robot interactions.

Learning by observation is an intuitive and interactive method used to train robots. To allow for incremental and extensible learning, a robot should be able to compose new knowledge form previously learned knowledge. In this paper, we aim to make a robot learn a human's complex gestures by combining primitive gestures that are predefined.

Nakamura et al. [4] has already tried to make a robot mimic a human's continuous motion via segmentation into predefined gestures. The applied gestures, however, are limited to repetitive and complete actions such as walking, running, and kicking. The method does not divide a continuous gesture into two single gestures because it does not find both starting and ending points correctly. In order to make a new motion from incomplete gestures, we consider that both starting and ending points should be recognized in human gestures that are not repetitive.

This paper presents a novel approach to segment continuous human gestures. First, a robot trains primitive gestures such as simple arm gestures through corresponding human gestures. When a trainer performs an action that he wants to teach to the robot, the robot segments the complex motion into its predefined primitive gestures. Second, when the complex motion is composed of ambiguous primitives that have a common area connecting each gesture, we make the robot find the optimal separating spot. Finally, the robot imitates the complex motion by following a human's demonstration.

We demonstrate an experiment where a small humanoid robot imitates a user's arm actions using three accelerometers. The proposed methods are applied to segment ambiguous actions and are evaluated with respect to various types of human gestures. From this experiment, the desired actions can be trained through human demonstration without direct intervention from the user.

2. Gesture Segmentation using DTW

In order for a robot to learn human gestures by demonstration, it needs to consider several characteristics of gestures. First, most gestures that humans perform are composed of a sequential combination of primitive gestures. Second, since a single gesture can be performed at various speeds the recognition algorithm should be able to deal with the temporal variation. Third, human gestures are performed smoothly even though the gesture is apparently divided into two separated gestures. In this section, we propose an approach to tackle how to segment the continuous gestures with temporal variation, and then we treat the third problem in the following section.

We introduce a solution that divides the user's motion, which is composed of more than two primitive gestures, into separate primitive gestures and subsequently recognizes each gesture. The robot recognizes the primitive gestures of users through accelerometers. We know the range of each primitive gesture through a collection of several data samples. The amount of time it takes for a user to complete a motion varies for every test, therefore, we apply the Dynamic Time Warping (DTW) algorithm. DTW is an algorithm for measuring similarity between two sequences that may vary in time or speed [5]. Since DTW can cope with different motioning speed, it has been applied to gesture segmentation [6,7].

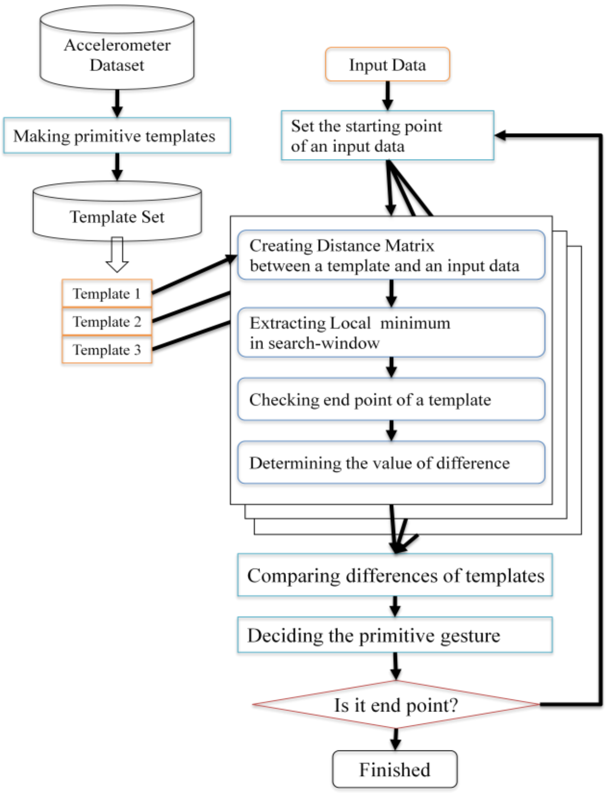
Accelerometer Dataset
Making primitive templates
Template Set
Template 1
Template 2
Template 3
Input Data
Set the starting point of an input data
Creating Distance Matrix between a template and an input data
Extracting Local minimum in search-window
Checking end point of a template
Determining the value of difference
Comparing differences of templates
Deciding the primitive gesture
Is it end point?
Finished

Fig. 1. The overall process of gesture segmentation using DTW. The gray box means the separated process according to the predefined gesture templates.

We describe the method for dividing and recognizing a continuous gesture into primitive gestures as shown in Fig. 1. A robot makes a template of each primitive gesture by repetitively capturing the user's gesture. The robot compares the user's input data with previously obtained primitive templates. The first primitive gesture in the input data is obtained by applying *dynamic programming* to every primitive template. The chosen primitive template is the one that has the lowest distance calculated by dynamic programming. The position that comes after the first primitive gesture in the input data is considered as a starting point of the second gesture. The robot tries to discover the second gesture using the same process. This process is iterated until user's data is completely processed.

Figure 2 illustrates the segmentation of a continuous gesture using the DTW algorithm. For example, there are three different primitive gestures and a user's input data consisted of the primitive gestures. The accumulated distances of three templates are respectively calculated based on the starting point of an input data. After calculation, the template that has the lowest distance at the last points is selected as a recognized gesture. The position after the end-point of the first gesture in the input data is considered as the starting point of next gesture. Then the second gesture is found using the same process. If the input data has no more data to compare the robot compares distances at the end of the input data for each primitive template, and then it selects the template with the lowest distance for the last gesture.

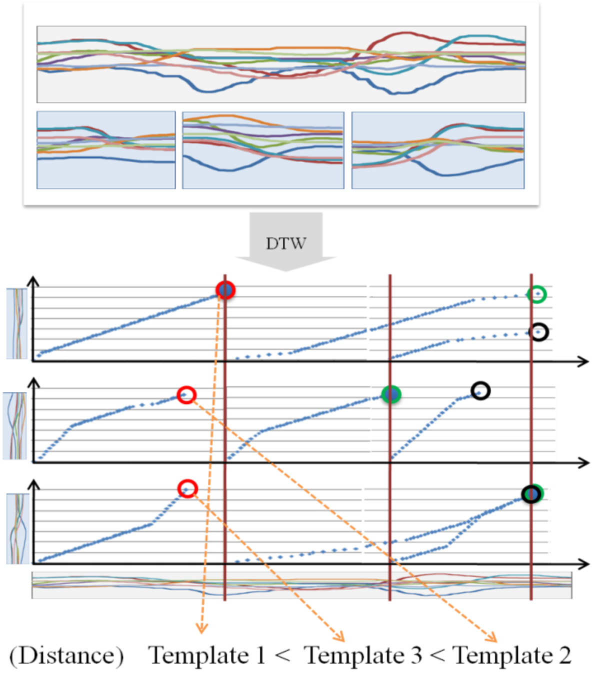

Fig. 2. Segmentation of a continuous gesture by DTW. Red, green, and black circles mean first, second, and third end-points respectively which were calculated by distance comparison.

3. Seamless Gesture Recognition

Continuous motion is not always composed of distinctly separated primitive gestures, as shown in Fig. 3. We define this motion as a *seamless gesture*. Since this gesture contains a *double-information* (e.g. end of the first primitive gesture and beginning of the second primitive gesture), finding accurate end-points is hard when segmenting input data using DTW. During segmentation, the calculated starting point of a second gesture may not be detected at the head but in the middle of the gesture. In this section, therefore, we propose a method that searches for an appropriate starting point in an area where double information has appeared.

When input data corresponds with a certain primitive gesture, we assume that the distributions of two data sets are similar. Furthermore, if some data area contains double information, the area could have different distribution values comparing with those of the corresponding template. After finding the end-point of the first gesture, both the recognized input data and the selected template are divided into several small blocks. Each block has its own distribution and we determine the similarity of distribution between the corresponding blocks.

In order to calculate the distribution we apply Kullback-Leibler (KL) divergence [8]. KL divergence of two data sets (P and Q) is represented as $D_{KL}(P||Q)$ and the equation is similar to the relative entropy (Eq. (1)).

$$D_{KL}(P \parallel Q) = -\sum_x p(x)\log q(x) + \sum_x p(x)\log p(x) \qquad (1)$$

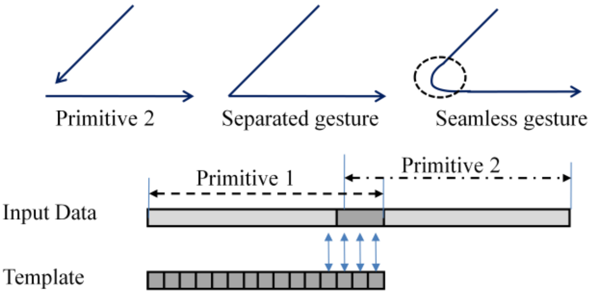

Fig. 3. The characteristic of seamless gestures. Both the dot circle of upper side and the dark-gray block of input data in the bottom side indicate ranges that contain double information.

where q(x) and p(x) denote the probabilities of data x in set of P and Q. If D_{KL} (P||Q)is zero, distributions of two sets are equal. The bigger the value is the more different two distributions are.

KL divergence is calculated for every small corresponding block between the template and the input data in order to compare distribution differences. If the distribution of data is different in the block, the input data is considered to have double information.

4. Experimental Result and Discussion

4.1 Experimental Result

In order to test our approach for gesture segmentation, we attached three 3-axis accelerometers on a user's arm, as shown in Fig. 4. A small humanoid robot is wirelessly controlled by a PC, which receives data from the accelerometers.

Before human demonstration, models of six primitive gestures are preset (e.g. up to down, left to right, up to left, left to down, down to right, and right to up). The training data of the user's gestures are previously acquired seven times, each at different speeds. The robot is able to perform each primitive gesture by predefined motion control program. At the start of the demonstration, a user shows several combined gestures, which are composed of primitive gestures. After the robot observes a human's demonstration, the robot analyzes the motion data and imitates the new combined gestures. Figure 7 shows some sequential gestures both that a human demonstrates and that a robot imitates after segmentation.

The experimental result is shown in Fig. 5 and 6. After recognizing continuous gestures using DTW, the accuracy rate of segmentation was about 60 to 80 percent on the users. While the length of combined gestures did not affect the accuracy rate remarkably, the accuracy was different depending on the testers. It is related to the existence of double-information in the gestures. Even if two testers perform the same gesture, one tester understands it as a connection of distinct primitive gestures and the other recognizes it as a seamless gesture. To evaluate that some seamless gestures with double information are critical in the continuous gestures, we applied KL divergence to the data. When applying only DTW, the accuracy rate was 71.9, 68.8, and 70 percent for sequential gestures. However, after adding KL divergence to search for optimal boundaries for the ambiguous gestures, accuracy was increased about 7 to 15 percent.

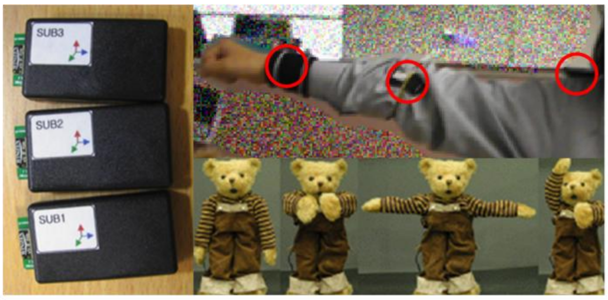

Fig. 4. Accuracy rates of gesture recognition according to testers and sequential gesture lengths

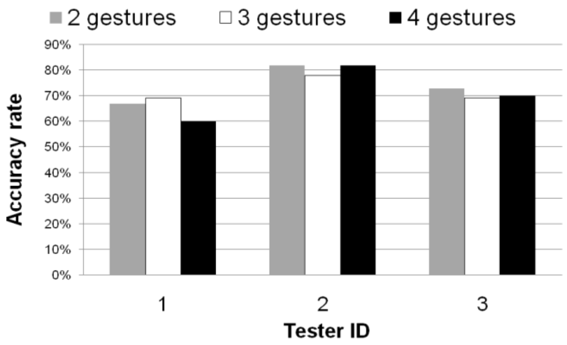

Fig. 5. Accuracy rates of gesture recognition according to testers and sequential gesture lengths

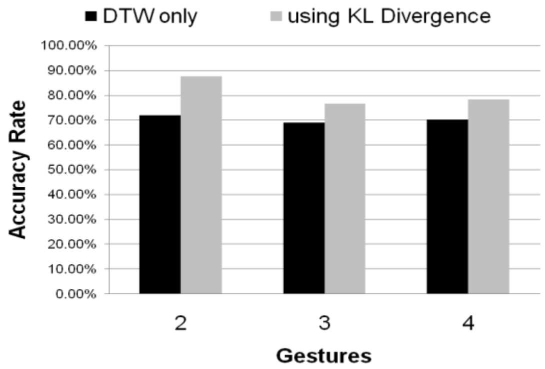

Fig. 6. Accuracy rates of gesture recognition according to gesture lengths and applied algorithms.

4.2 Discussion

We found that several additional works enable robots to be more natural during human-robot interaction, in addition to imitation learning. First, a general template is necessary for a robot to recognize gestures regardless of the identity of users. Second, a robot needs to extend its performable gestures such as hand, leg, and body motions by imitation. One method is to include discriminative primitive gestures based on other motion sensors and a vision system. On the other hand, our experiments only evaluated partial gestures generated by three

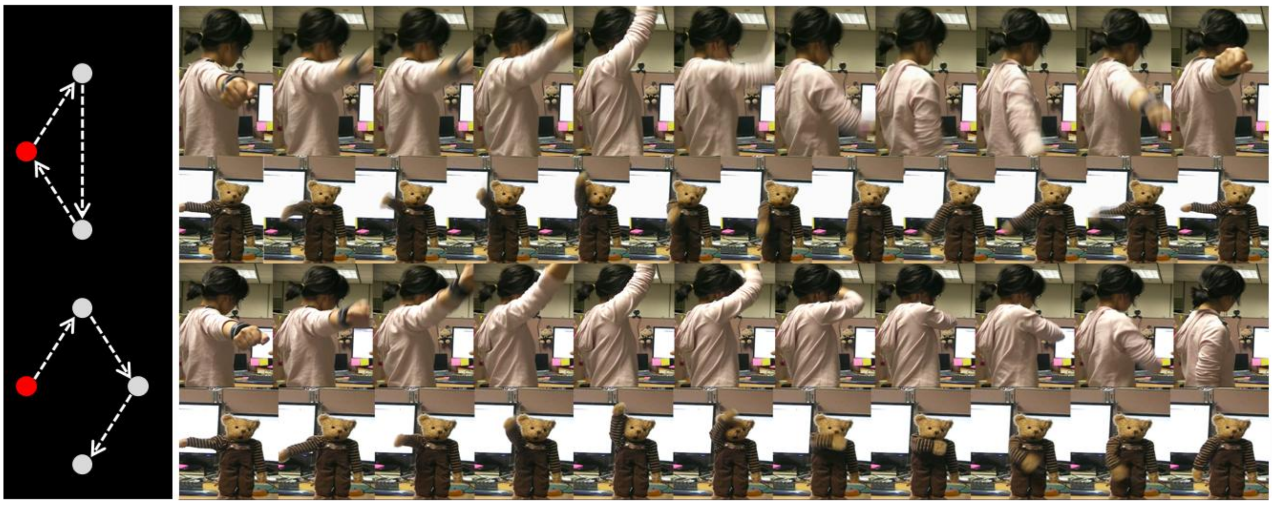

Fig. 7. Human demonstration with accelerometers for robot imitative learning. The red circle means the starting position of a continuous gesture.

accelerometers on an arm. Another method is to account for the speed and levels of smoothness in doing imitative gestures. A single gesture may have various meanings depending on the intentions of the user. The intention usually affects some factors such as speed, power, and route. If a robot is aware of the details of gestures and reflects them to imitate a motion, combination of same primitive gestures can be extended to various types that have their own meaning. Fourth, we give feedback to a robot by applying other interfaces like voice and display, as a method for evaluation of learning. These interfaces allow a robot to learn human gestures more actively and interactively.

5. Conclusion

In this paper, we proposed an algorithm to segment complex/continuous human motions into a combination of primitive gestures that are previously defined and known to a robot. Using DTW, a sequentially connected gesture was divided into predefined elemental actions. Moreover, when two gestures were merged into one seamless gesture, we found the exact starting point of the following gesture by using KL divergence. By combining the two approaches, we achieved about 80 percent accuracy of gesture recognition.

Imitation learning is an important factor in human-robot interaction to increase the applicability of robots. This paper aimed to solve the fundamental problem of imitation and to extend gestures that a robot can perform without predefined models. This research can be a basis for interactive and imitative learning.

References

[1] H. Kang, C. W. Lee and K. Jung, "Recognition-Based Gesture Spotting in Video Games", Pattern Recognition Letters 25, pp.1701-1714, 2004.

[2] J. Alon, "Spatiotemporal Gesture Segmentation", Boston University. 2004.

[3] T. Inamura, N. Kojo and M. Inaba, "Situation Recognition and Behavior Induction based on Geometric Symbol Representation of Multimodal Sensorimotor Patterns", In Proc. of ICIRS, 2006.

[4] W. Takano and Y. Nakamura, “Humanoid Robot's Autonomous Acquisition of Proto-Symbols through Motion Segmentation”, IEEE-RAS International Conference on Humanoid Robotics, pp. 425-431, 2006.

[5] H. Sakoe, and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition”, IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1) pp. 43- 49, 1978.

[6] M. H. Ko, G. West, S. Venkatesh and M. Kumar, "Online Context Recognition in Multisensor Systems using Dynamic Time Warping", ISSNIP, pp.283-288, 2005.

[7] H. Li and M. Greenspan, “Segmentation and Recognition of Continuous Gestures”, ICIP, 2007.

[8] S. Kullback, “The Kullback-Leibler distance”, The American Statistician 41:340–341, 1987.