



Evaluating Instructions for Gesture Recognition with an Accelerometer

Murao, Kazuya

Terada, Tsutomu

(Citation)

ヒューマンインタフェース学会論文誌, 17(1):73-84

(Issue Date)

2015-02-25

(Resource Type)

journal article

(Version)

Version of Record

(Rights)

© 2015 ヒューマンインタフェース学会

本著作物の著作権は特定非営利活動法人ヒューマンインタフェース学会に帰属します。本著作物は著作権者であるヒューマンインタフェース学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」に従うことをお願い致します。

(URL)

<https://hdl.handle.net/20.500.14094/0100492901>



Evaluating Instructions for Gesture Recognition with an Accelerometer

Kazuya Murao^{*1} and Tsutomu Terada^{*2 *3}

Abstract – In the area of activity recognition with mobile sensors, a lot of works on context-aware systems using accelerometers have been proposed. Especially, mobile phones or remotes for video games using gesture recognition technologies enable easy and intuitive operations such as scrolling browser and drawing objects. Gesture input has an advantage of rich expressive power over the conventional interfaces, but it is difficult to share the gesture motion with other people through writing or verbally. Assuming that a commercial product using gestures is released, the developers make an instruction manual and tutorial expressing the gestures in text, figures, or videos. Then an end-user reads the instructions, imagines the gesture, then perform it. In this paper, we evaluate how user gestures change according to the types of the instruction. We obtained acceleration data for 10 kinds of gestures instructed through three types of texts, figures, and videos, totalling 44 patterns from 13 test subjects, for a total of 2,630 data samples. From the evaluation, gestures are correctly performed in the order of text→figure→video. Detailed instruction in texts is equivalent to that in figures. However, some words reflecting gestures disordered the users' gestures since they could call multiple images to user's mind.

Keywords : Gesture recognition, Instructing gestures, Accelerometer

1. Introduction

Along with the downsizing of computers, many devices containing small sensors such as accelerometers have been released. In particular, an accelerometer is installed in most current mobile phones, such as the iPhone and Android-powered devices, and video game controllers for Wii and PS3, which enables easy and intuitive operations such as scrolling browser and drawing graphical object that have been difficult to input with conventional interfaces^[9].

In general, a gesture recognition system constructs a model with data and ground truth captured beforehand, then recognizes gestures that end-user performed by comparing with the model. Therefore, one problem arises when the system is distributed or on sale. System developer cannot directly instruct end-users on gestures, but need to convey through manual or tutorial. In detail, system developers make a model of gestures to invoke some commands they assume. The data of the gestures are captured on the developer's side. Then, they make an instruction manual expressing the gestures in texts, figures,

or videos, which is delivered to the end-users. The end-users read the instruction and imagine gestures then perform them. However, the end-users may imagine different gestures to ones developers assume. Although the developers assume pitching form of a baseball pitcher and express it as “throw”, the end-users may imagine throwing a Frisbee, throwing a trash underhand, or tossing a coin. This problem would be caused by several reasons: system developer's expression is distorted or ambiguous, or originally it is hard to convey gestures because gesture has rich expressive power.

A lot of studies on gesture recognition have been done thus far. Seeing the evaluations in these, the ways of collecting data are diverse. Though training data and test data are collected from same person or different person, or cross-validation is used, how gestures have been conveyed are not clearly described in these papers^{[5], [7], [9], [12]}. It, however, is suspected that the test subjects could directly and orally be instructed to do correct gestures since using incorrect gestures would cause unreliable evaluation and result. Also many studies on gesture learning and instructions have been proposed^{[1]~[4]}. Though these works have proposed systems which enable users to input correct gestures and investigated appropriate mapping of gestures to commands for mobile devices,

^{*1}: College of Information Science and Engineering, Ritsumeikan University, Japan

^{*2}: Graduate School of Engineering, Kobe University, Japan

^{*3}: PRESTO, Japan Science and Technology Agency, Japan

they focus on 2D gestures on touch screen or mouse gestures.

This paper evaluates how users' gestures change according to the types of the instructions. Gesture movements are captured with accelerometer mounted on user's wrist and instructions of gestures are expressed in three types of texts, in figure, and in video.

2. Related work

2.1 Gesture recognition

Graeme et al. proposed a video annotation method by gesture recognition using one accelerometer mounted on wrist since video annotation is difficult only from analyzing video^[5]. They use Hidden Markov Model (HMM)^[11] for recognition, resulting in only one mistake for 30 trials of three Kung-Fu martial art movements; Cuts, Elbow, and Punch. In this experiment, data is obtained from only one subject. Since training and testing are carried out with data of same person, there is no room for miscommunication.

Georgia Tech Gesture Toolkit^[12] as a tool to support gesture recognition has been proposed by Westeyn et al. This is a toolkit that enables ordinary users who do not have enough knowledge on speech recognition to use existing HMMtoolkit^[6] with ease. In the literature, four applications are presented, one is gesture recognition with two 3-axis accelerometers positioned at wrist and elbow, and achieves 93.3% accuracy for ten kinds of gestures such as grinding, sawing, and screwing. Though each gesture is ambiguous, also this experiment captures the data from one subject for ten times. Same gestures would be iterated through his/her task scenario.

The method proposed by Liu et al. recognizes eight gestures such as drawing a line and a circle, which are recommended by Nokia laboratory, with one 3-axis accelerometer^[9]. This research captures more than 4,000 samples from eight test subjects for a long period. They use Dynamic Time Warping algorithm (DTW)^[10] as a recognition algorithm and achieve 98.6% accuracy by successively renewing training data. These eight gestures are expressed with arrows as the paths of hand movements, identified by users for interaction with home appliances. Therefore, they are simple and can be expressed with arrow. However, gestures that cannot be expressed

with figures are not discussed, besides how correctly the test subjects performed the gestures is not mentioned. In addition, data of same person are used for training and testing since this work investigates the change in the individual gesture movements day by day.

The method proposed by Junker et al. recognizes ten daily short actions such as push button, drink, and phone up/down, and achieved approximately 80% precision and recall^[7]. Innovative point of the study is partitioning sensor stream into several segments that represent each human movement by the sliding-window and bottom-up (SWAB) algorithm^[8]. In the experiment of the paper, data sets are recorded from one female and three male right-handed subjects, aged 25-35 years. The purpose of the study is explained to the subjects. However, the subjects were asked to perform the movements as natural as possible while wearing the sensors. The examiners did not set constraints to the movements of the subjects, except that they asked the subject to perform the relevant gestures according to the description given in the table in the paper. Examples of the description are *"Take out purse from right back pocket of trousers. Open purse with right hand. Take coin and insert it into slot of vending machine. Close purse with right hand. Put purse back into pocket"* as coin gesture and *"Meal intake of Lasagne using fork and knife. Fork tap, loading and manoeuvring to mouth and back with left hand"* as cutlery gesture. Though these descriptions are detailed and subjects would correctly perform gestures, how many subjects misunderstood the descriptions and how do they work in different description are not mentioned.

2.2 Gesture learning

Alexander et al. examined a mapping of foot-based gestures to mobile device commands and probed continuous interaction gestures^[1]. In this literature, the experimenter placed the command into context, for example, *"Your phone is ringing and you want to answer it. What foot gesture would you perform to answer the phone?"* and the participant then performed a gesture they believed appropriate for this task. They were able to define gestures for 23 of 30 commands in the command set including "Answer incoming call" and "Home screen" for phone control, "Volume up" for media control, and "Zoom in" for map navigation. They also

probed continuous foot interaction such as adjusting the volume of the audio and proposed four gesture mapping techniques: displacement based, rate-based hold, rate-based continuous, and flick. OctoPocus proposed by Bau et al. is a dynamic guide that combines on-screen feedforward and feedback to help users learn, execute and remember stroke gesture commands on touch screen^[2]. OctoPocus continuously updates the state of the recognition algorithm by gradually modifying the thickness of possible gesture paths, based on its error rate. In these studies, gestures are not strongly associated with commands and freely configurable, therefore the users can map any gestures to the command so that the users easily guess, perform, and remember. However, gestures in video game should be same actions of a character in the game, e.g. throwing gesture cannot be mapped to opening door gesture.

Bragdon et al. proposed GestureBar, a user-interface for learning gestural interactions that enables a walk-up-and-use experience which is in the same class as standard menu and toolbox bar interfaces^[3]. GestureBar has a look of a common toolbar as in Microsoft Office, but in place of executing commands, discloses how to execute commands with gestures through animated images, detail tips and an out-of-document practice area. This tool is for lineogrammer gesture set in 2D space, which can be displayed with lines and dots, but gesture motions in 3D space are hardly depicted in image and animation.

Fothergill et al. investigate the question of what is the most appropriate semiotic modality of instructions for conveying to human subjects the movements the system developer needs them to perform^[4]. They investigated three different semiotic modalities: Descriptive text, static images, and dynamic video, as well as combination of images with text and video with text, and showed that different semiotic modalities of an instruction alter the amount of variation in a set of training performance. The number of types of text is one in this work, but we investigate the effect of difference in expression of text on user gestures and confirmed that detailed text is not simply effective for collecting data. This literature concluded that instructions in video or video+text performed well. Though it is better for users to watch video or to play tutorial at an

easy pace, it would be hard to retrieve gesture command during playing video game or when abruptly inputting command since video has low browsability. Investigation is needed on expressions in text that can be understood at a glance and can be played in audio.

3. Experimentation

3.1 Setup

The gesture data were taken from ten male and three female subjects aged 21 to 24 years, who wore a 3-axis wireless accelerometer^[13] on their right wrist. The sampling frequency was 50 Hz. They acted ten kinds of gestures that could be performed in video games:

- Draw a circle: Subjects draw a circle counterclockwise starting from the top in their front.
- Draw a triangle: Subjects draw a triangle counterclockwise starting from the top in their front.
- Draw a square: Subjects draw a square counterclockwise starting from upper left in their front.
- Draw a star: Subjects draw a star starting from top to lower left.
- Punch: Subjects punch straight forward.
- Push door: Subjects move their arm as though pushing and opening a door.
- Pull drawer: Subjects move their arm as though pulling and opening a drawer.
- Pull door: Subjects move their arm as though pulling and opening a door.
- Throw: Subjects move their arm as though throwing a ball forward like a baseball pitcher.
- Chop: Subjects chop at a slant in their front.

by following the instructions expressed in three types of texts, in a figure, and in a video. The descriptions of the gestures above are not explained to the subject. The instructions in text are listed in Table 1. All the subjects are Japanese and the instructions were given in Japanese, therefore there is no linguistic misunderstandings.

Instruction no. 1 is vague and no. 3 is detailed. Push door, pull door, and throw gestures instructed through no. 1 were not captured since these instructions are same as those of punch, pull drawer, and chop, respectively.

Instructions of circle, triangle, square, and star were given in figures shown in Figure 1, in which dot denotes the starting point and arrow denotes

Gesture	No.	Instruction
Circle	1	Draw a circle
	2	Draw a circle from top
	3	Draw a circle counterclockwise from top
Triangle	1	Draw a triangle
	2	Draw a triangle from top
	3	Draw a triangle counterclockwise from top
Square	1	Draw a square
	2	Draw a square from upper left
	3	Draw a square counterclockwise from upper left
Star	1	Draw a star
	2	Draw a star from top
	3	Draw a star from top to lower left
Punch	1	Put your hand forward
	2	Stick your hand forward
	3	Punch
Push door	1	Put your hand forward
	2	Push your hand forward
	3	Push your hand forward like opening a door
Pull drawer	1	Pull your hand horizontally
	2	Pull your hand horizontally with your palm upwards
	3	Pull your hand horizontally like opening a drawer
Pull door	1	Pull your hand horizontally
	2	Pull your hand horizontally with your palm downwards
	3	Pull your hand horizontally like opening a door
Throw	1	Swing your arm
	2	Throw
	3	Throw a ball
Chop	1	Swing your arm
	2	Chop
	3	Chop at a slant

Table 1 Instructions in text.

the direction, and instructions of other gestures in figures were not carried out since there are no figures expressing these gestures in common. Video of gestures are moving image of gestures performed by one of the authors.

3.2 Data collection

Acceleration data were captured through our application shown in Figure 2, which displays texts in Table 1, figures in Figure 1, videos, and waveform of acceleration. The test subjects imagined one gesture from each instruction and performed it five times. The subjects were asked to perform the gestures as same as possible for one instruction. The experiment was video-recorded for later verification.

At first, the instructions of the ten kinds of gestures in text 1 were given to the subjects. Order of the gestures was randomized. After all the instructions in text 1 have finished, instructions in text 2 were given. Order of the gestures was randomized as well. Then, instructions in text 3, figure, and video were given in this order in the same manner. Instructions with same instruction type were given from the vague ones so that the subjects cannot infer correct

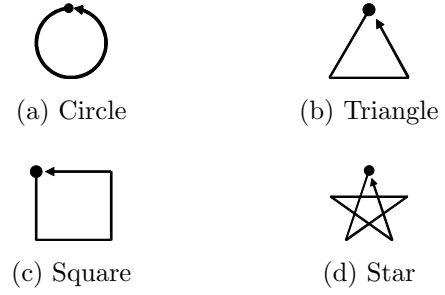


Fig. 1 Figures of gestures.

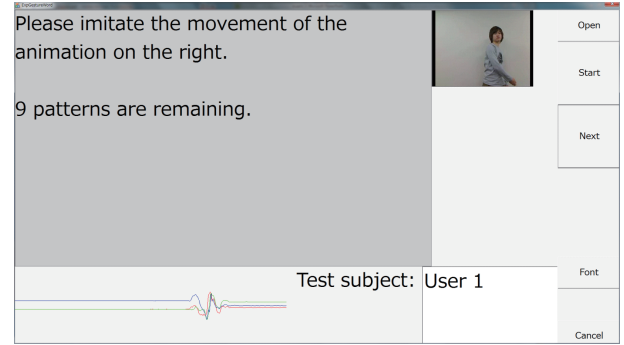


Fig. 2 Application used in the data collection.

gestures from the detailed instructions. Experiment finishes after all the instructions in video were done.

The subjects were asked to stand straight between each gesture to separate the data into gesture and non-gesture with ease and to remove the effect by the hand position and orientation before and after gesturing. By holding these manners, only the difference in gesture movements can be focused on. Data of 44 patterns from 13 subjects were collected through the experiment. Since several samples were not collected due to transmission errors, 2,630 samples were collected in total. At least four samples were collected for each instruction.

3.3 Similarity calculation algorithm

Time-series data has been used in various fields such as science, medicine, economics, and engineering. Calculation of similarity between time-series data is required in order to do data-mining in these fields. Though the simple method to measure similarity is Euclidean distance, it has drawbacks that it is susceptible to temporal distortion and numbers of samples of two data sequences must be equal.

Dynamic Time Warping (DTW)^[10] is an algorithm to measure the similarity of two time-series data, which redeems the drawbacks of Euclidean distance. DTW calculates a temporal non-linear elastic

distance between two sequences, which may vary in time, and the number of both samples need not be equal. For example, comparing two kinds of data for draw a circle in the air whose rotating speed are different, DTW can find their similarity. In addition, in case that a part of each data differs, DTW is applicable because of its nonlinear elasticity.

Recognition systems of hand-written characters often use their trajectories by calculating double integral of acceleration values since the trajectory is important. Gestures, however, should be discriminated also by their speed and intensity even if their trajectories are similar e.g, punch vs. push and slap vs. swipe. Raw sensor values are generally processed as time-series for such gestures.

Hidden Markov Model (HMM) [11] is also famous for a method to treat time-series data and some studies use HMM to recognize gestures. However, HMM outputs likelihood of an unknown gesture to be a predefined gesture, while DTW measures spatial distance between gestures. Since difference in distance is better than that in likelihood to measure the difference between time-series, we adopt DTW in this paper.

The algorithm works as follows. For the sake of simplicity, we assume there is data for one axis. Time-series data $X = (x_1, \dots, x_m)$ and $Y = (y_1, \dots, y_n)$, with length m and n , respectively, are compared. An $m \times n$ matrix d is defined: $d(x_i, y_j) = |x_i - y_j|$. Next, warping path $W = (w_1, \dots, w_k)$, which is a path of pairs of indices of X and Y , is found. W meets three conditions.

- Boundary
 $w_1 = (1, 1), w_k = (m, n)$
- Seriality
 $w_k = (a, b), w_{k-1} = (a', b')$
 $\Rightarrow a - a' \leq 1 \wedge b - b' \leq 1$
- Monotony
 $w_k = (a, b), w_{k-1} = (a', b')$
 $\Rightarrow a - a' \geq 0 \wedge b - b' \geq 0$

To find the path with the lowest cost that meets these conditions, the following steps are applied.

1. Initialization
 $DTW(0, 0) = 0$
 $DTW(i, 0) = \infty$ for $i = 1, \dots, m$
 $DTW(0, j) = \infty$ for $j = 1, \dots, n$
2. Do for $i = 1$ to m
 Do for $j = 1$ to n

$$DTW(i, j) = d(x_i, y_j) +$$

$$\min \begin{cases} DTW(i-1, j-1) \\ DTW(i-1, j) \\ DTW(i, j-1) \end{cases}$$

3. Output

Return $DTW(m, n)/(m + n)$

The obtained DTW distance becomes a distance between two time-series of gestures. Value of DTW distance shows the degree of difference between two gestures, therefore small distance means the two gestures are similar. The returned DTW distance is divided by sum on length of the time-series since DTW distance increases in response to the length of the two sequences. The original DTW calculation is configured for one dimensional data sequence, applying for 3-axis accelerometer, DTW calculation is carried out for each axis and sum of the distance for three axes is used as a distance of gestures.

4. Evaluation

4.1 Difference in motion by visual check

This section discusses the difference in motion by checking the video that has recorded subjects' motions during the experiment. Table 2 shows the number of subjects who performed correct gestures and the incorrect motions performed for each instruction. The parenthetic values following the description of incorrect gestures are the number of subjects who performed the gesture. Incorrect gestures for drawing figure gestures are illustrated in figures, in which dot denotes the starting point and arrow denotes the direction. The numbers of subjects for "push door", "pull door", and "chop" gestures instructed in text 1 are represented by '-' marks since instructions of these gestures were same as those of "punch", "pull drawer", and "throw" gestures, respectively. Also "punch", "push door", "pull drawer", "pull door", "throw", and "chop" gestures instructed in figure are represented by '-' marks since these gestures were not captured.

From the results of circle gesture in text 1, nine test subjects drew a circle from the bottom, all of which are clockwise. However, five subjects changed the drawing direction to counterclockwise just by aligning the starting point to top in text 2. All the subjects drew a circle in their front and no subjects drew a circle in their right side like pitching a softball. For

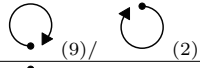
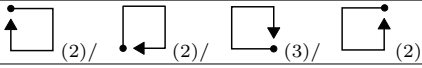
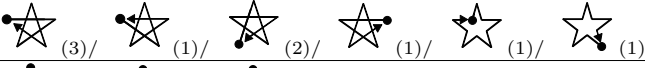
Gesture	Type	# subjects who gestured correctly (out of 13)	Incorrect gestures and the number of subject who performed the gesture
Circle	Text 1	2	 (9)/ (2)
	Text 2	9	 (4)
	Text 3	13	
	Figure	13	
	Video	13	
Triangle	Text 1	9	 (3)/ Correct triangle but position of wrist is high(1)
	Text 2	12	Correct triangle but position of wrist is high(1)
	Text 3	12	Correct triangle but position of wrist is high(1)
	Figure	13	
	Video	13	
Square	Text 1	4	 (2)/ (2)/ (3)/ (2)
	Text 2	8	 (5)
	Text 3	12	 (1)
	Figure	13	
	Video	13	
Star	Text 1	4	 (3)/ (1)/ (2)/ (1)/ (1)/ (1)
	Text 2	10	 (1)/ (1)/ (1)
	Text 3	11	 (2)
	Figure	13	
	Video	13	
Punch	Text 1	2	Very slow punch(2) / Put hand up flatly(1) / Swing arm up like walking(6) / Push hand forward with palm heading for far side(1) / Put hand up flatly with palm facing down(1)
	Text 2	5	Very slow punch(1) / Put hand up flatly(3) / Swing arm up like walking(3) / Karate-punch; Punch quickly and pull it back quickly(1)
	Text 3	11	Karate-punch; Punch quickly and pull it back quickly(1) / Twisted punch(1)
	Figure	-	
	Video	13	
Push door	Text 1	-	
	Text 2	1	Push hand forward with palm heading for far side(11) / Push hand forward with palm heading down(1)
	Text 3	8	Push hand forward with palm heading for far side(3) / Pull door with turning a doorknob(1) / Pull door with turning a door handle(1)
	Figure	-	
	Video	13	
Pull drawer	Text 1	4	Collect pull drawer but the position of the hand is high(1) / Pull with thumb heading up(1) / Pull with thumb heading left (similar to pull door gesture)(6) / Pull like pulling a fishing rod(1)
	Text 2	13	
	Text 3	11	Pull drawer with pinching a tab(2)
	Figure	-	
	Video	13	
Pull door	Text 1	-	
	Text 2	11	Collect pull door but the position of the hand is high(2)
	Text 3	9	Pull with thumb heading up(3) / Pull door with turning a doorknob(1)
	Figure	-	
	Video	13	
Throw	Text 1	0	Swing arm like knocking a door three times(2) / Swing arm right to left like a pendulum(1) / Swing arm up like walking(3) / Swing arm up and back like walking(2) / Swing arm up, back and up like walking(2) / Swing arm up, back, up and back like walking(3)
	Text 2	12	Throw to left(1)
	Text 3	10	Throw underhand(2) / Throw to left(1)
	Figure	-	
	Video	13	
Chop	Text 1	-	
	Text 2	3	Down-stroke chop(10)
	Text 3	13	
	Figure	-	
	Video	13	

Table 2 Difference in motions by visual check.

triangle gesture, nine subjects drew collect triangles without instructions of drawing direction and starting point in text 1. Three subjects drew it from

the lower left. All the subjects drew it in counter-clockwise. No one drew a triangle upside down like ∇ . Drawings of square gesture are diverse and their

directions and starting points are correct only after these instructions are given in text 3. One subject drew it clockwise though instructed counterclockwise in text 3 since the subject misread the instruction. Complex results came up for star gesture. Four subjects gestured correctly in text 1. Interestingly, two subjects imagined not a five-stroke star as in Figure 1 but an outer frame of a star. One of them got aware that it has been incorrect after the figure is shown.

The other six gestures than drawing a figure gestures lead to more complex results. Six subjects swung their arm up like walking for the instruction of punch for text 1. For text 2, though the number of subject who gestured correctly increased from two to five, the way of incorrect gestures is still diverse. Giving the instruction expressing simply “punch”, eleven subjects gestured correctly. Instruction of push door gesture in text 2, which does not include a word “door”, called eleven subjects pushing horizontally with their palms heading for far side. Though the instruction for text 3, which includes a word “door”, improved the number of subject who did correct gestures from one to eight, it called subjects turning a doorknob and turning door handle.

Seeing the motions when the instruction of “pull” for text 1 is given, they look similar but palms are heading for different direction. Therefore, indicating the direction of palm, all subject answered correctly for both pull drawer and pull door. Two subjects raised their hand to the height of the eye. Instruction for text 3 that includes a word “drawer” make two subjects call a tab and they pinched it. A word “door” invoked a subject to turn a doorknob and three subjects to open a door with their palms facing to left as if opening a fridge door.

The instruction of throw gesture in text 1 was so ambiguous that no subject imagined throw nor chop, two of them knocked three times and ten of them swung arm like walking or a pendulum. Given a word “throw” in text 2, all subjects threw but one of the subjects threw left as if a baseball pitcher picks off the first base. Moreover, adding a word “ball”, two subjects changed their forms to throwing underhand even though they gestured correctly when text 2 was given. Ten subjects performed down-stroke chop and three subjects correctly performed chop on the slant just by expressing chop as “chop”. No one

chopped horizontally like a wrestler. Adding a word of “on the slant”, all the subjects gestured correctly.

For the instructions in figure and video, all the subjects performed correctly, though they had slightly different motions.

4.2 Difference in gestures by similarity calculation

This section quantitatively evaluates how much the gestures of test subjects differ from correct one by measuring DTW distance. Specifically, given data sequences of circle gesture in text 1 for five times ($1_a, 1_b, 1_c, 1_d, 1_e$), each of them is compared with data of the same subject collected through the video instruction for five times (v_a, v_b, v_c, v_d, v_e). Let distance of 1_a be $dist(1_a) = \min\{D(1_a, v_a), D(1_a, v_b), D(1_a, v_c), D(1_a, v_d), D(1_a, v_e)\}$, where $D(x, y)$ is a DTW distance between time-series x and y . Distances of $1_b, 1_c, 1_d$, and 1_e are calculated in the same way, and distance of circle gesture for text 1 is given by average of $dist(1_a), dist(1_b), dist(1_c), dist(1_d)$, and $dist(1_e)$. For the data obtained through video instructions, distance of v_a is exceptionally calculated between v_a and the other four samples as in $dist(v_a) = \min\{D(v_a, v_b), D(v_a, v_c), D(v_a, v_d), D(v_a, v_e)\}$ since $D(v_a, v_a)$ becomes zero. Distance of circle gesture for video is given by average of $dist(v_a), dist(v_b), dist(v_c), dist(v_d)$, and $dist(v_e)$. The small distance means that the gesture is similar to the correct one and the instruction expressed the gesture well.

Figure 3 shows mean and standard deviation for each subject and their average denoted by “all”. For the results, as an overall trend, DTW distances of drawing a figure gestures are shortened in the order of text 1→text 2→text 3→figure→video. Firstly, the standard deviations are almost same for the subjects and instructions except for few cases e.g. square gesture of subject 4 in text 2, and pull drawer gesture of subject 7 in text 3, since the subjects were asked to perform the gestures as same as possible for one instruction. Results of text 1 show that distance is large and the subjects were not instructed the gesture correctly, which accords with the result in the visual check. Seeing the results of text 3 and figure, the direction and starting point has equivalent communication competence to figure since these distance are close.

Instruction in video is equivalent to an instruc-

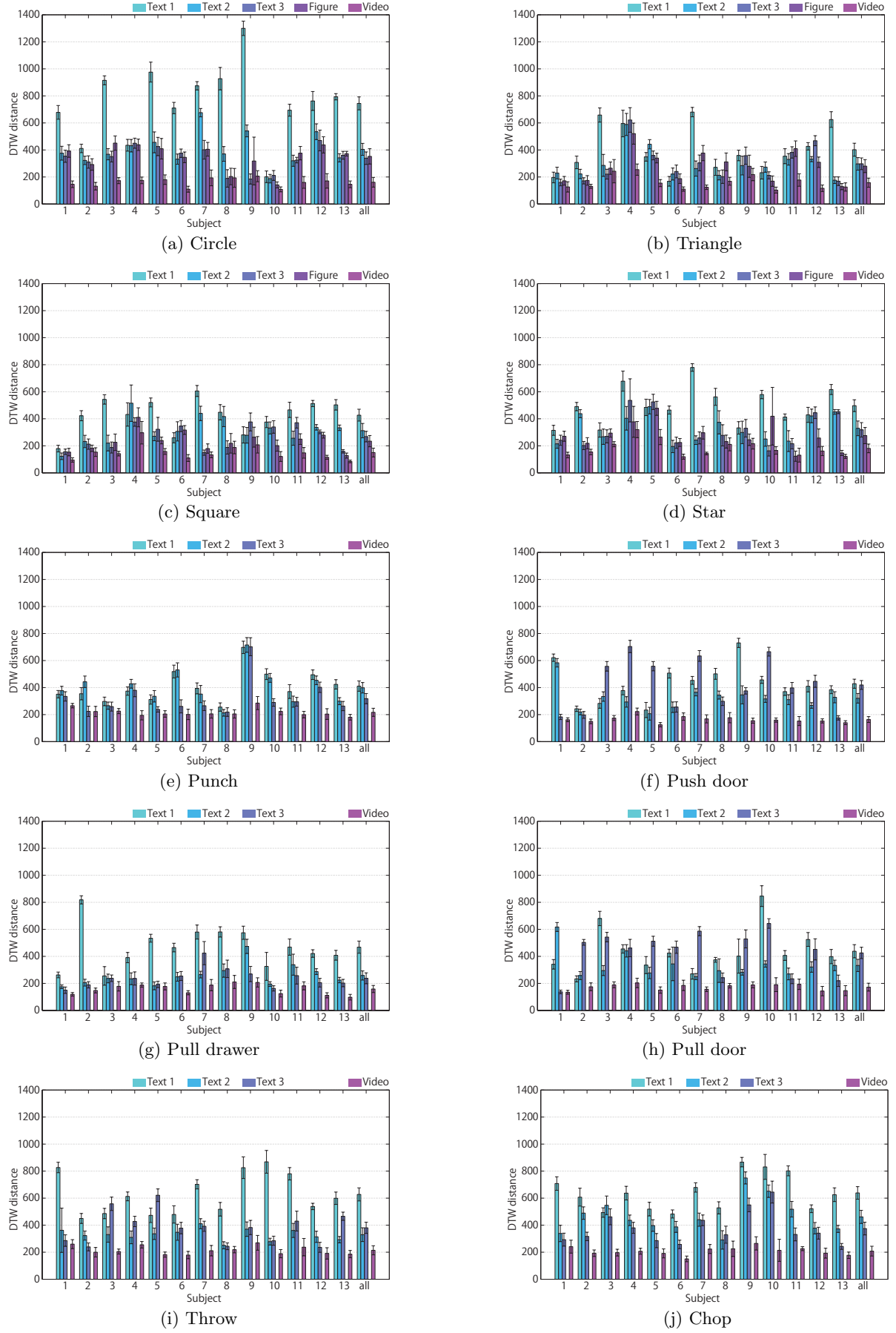


Fig. 3 DTW distance between each instruction and video.

tion showing an actual movement face to face. Even if showing the actual movement face to face, distance never become zero since reproducing the gestures completely is impossible. The distance of video instruction is the lower limit of the distance. Comparing the result of figure with the result of video, result of video is much smaller than that of figure, which means figures are inferior to videos.

DTW distance of the remaining six gestures also wholly decrease in the order of text 1→text 2→text 3→video. However, the distances of “push door” and “pull door” gestures instructed in text 3 are larger than those in text 2. This is because the word “door” hardly called the correct gestures to the subjects’ mind in common since there are a lot of types of door such as with doorknob, with door handle, without knob, and with a pole installed on high-class boutique. Therefore, the word “door” instead of the direction of palm made the instruction ambiguous.

On the other hand, such tendency did not appear for “pull drawer” gesture since most of all subject imagined a drawer which opens by pulling with inserting the hand to the bottom of it. However, since two subjects imagined a tab to be pulled, the word of an object that would call multiple types is not appropriate even though it is a keyword of the gesture.

For the result of punch, the result of text 2 shows nearly equal distance to the result of video by just indicating “punch”. Throw and chop instructed in text 2, however, did not mark small distance though they are one-word verbs expressing the gestures like punch. Since chop gesture has several directions, subjects did down-stroke chop that is different to the correct one when instructed in text 2. For these results, how much degree of freedom a gesture has should be discussed when it is defined.

Also “throw” gesture has several options such as overhand, sidearm, underhand, and Frisbee. Most of the subjects, however, threw overhand when instructed in text 2. In contrast, by adding the word “ball” to make it more detailed, two subjects originally threw overhand threw underhand, resulting in the longer distance. It also disordered the instruction like “door”.

4.3 Interpersonal diversity

This section evaluates the diversity among test subjects. Large diversity means that the instruction was vague and did not call common awareness to

their mind. Reversely, small diversity means that the instruction called common awareness, and the words were appropriate for instruction manuals. In that case, even though the awareness was not correct, the expression was meaningful and it would be said that the person who made the instruction was wrong.

Diversity is measured by calculating DTW distance among all the samples of same gesture and same type for all the subjects. Specifically, DTW distance of all the combinations between the data of one subject for circle gesture instructed in text 1 and data of the other subjects for circle gesture instructed in text 1 are measured, whose average becomes a diversity of the subject for circle gesture instructed in text 1. After this calculation takes place for the other subjects in the same way, averaged diversity over all the subjects becomes the diversity of circle gesture instructed in text 1. Diversity for the other gestures and instruction types are calculated in the same way. Diversities for each gesture and instruction are shown in Figure 4.

As a general trend, DTW distance decreases in the order of text 1→text 2→text 3→figure→video, which means common gestures were called. However, seeing the result of chop instructed in text 2, only the expression “chop” called incorrect gesture in common. In addition, as stated above, more diverse gestures were imagined from “push door”, “pull door”, and “throw” gestures instructed in text 3. This is because the words which had been added in order to make the instructions more detailed were translated in diverse way. These results indicate that adding words does not always cause better results. If the gesture is a motion using a object, it must be described in detail or detailed hand movement must be instructed.

4.4 Recognition accuracy

At last, this section shows the results of gesture recognition. The samples collected through instructions in video are used for training data. DTW distance between one sample and the training data are calculated, then the gesture of training data that produced the smallest distance becomes a recognition result. The results are shown in the confusion matrices in Table 3.

Each cell indicates the number of positives per activity (with the true positives diagonally), the recall

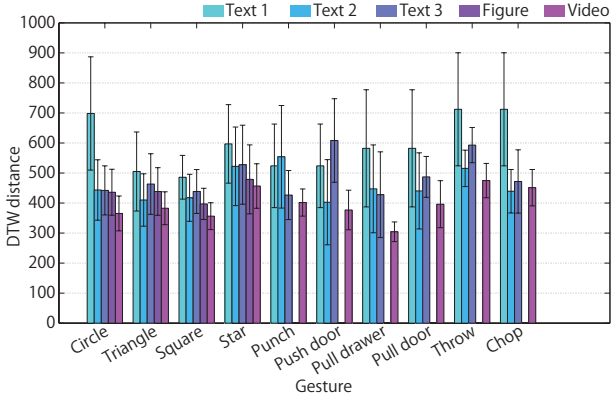


Fig. 4 DTW distance among subjects.

indicates the percentage of true positives over the input data samples for each gesture. The result shows that the recalls increase in the order of text 1→text 2→text 3→figure→video, which means that detailed instructions call the common gestures in the subjects' mind.

Seeing the results for each gesture in detail, gestures drawing a figure are missed each other, which is, however, not a critical issue since these gestures are recognized in approximately 0.96 accuracy when instructed in figures. Comparing the results of text 3 and figure, detailed instruction in text is equivalent to figure.

On the other hand, for the remaining six gestures, "pull drawer" and "pull door", and "throw" and "chop" gestures are recognized well each other when instructed in text 2 even though they are similar. Recalls of "punch" and "push door" gestures, which involve similar arm action and same palm orientation, instructed in text 2 are quite low and they are low even when instructed in text 3. However, they are recognized well when instructed through video, which means that they are discriminative if correctly performed and it is hard to express their difference in texts. Therefore, such gestures should not be used together.

In addition, if the training data does not include data of the user, the recall drops to 0.364 for text 1, to 0.754 for text 2, to 0.784 for text 3, to 0.957 for figure, and to 0.892 for video, respectively. Though recalls dropped in all types of instructions compared with the case the model is calibrated, similar trend appeared regardless of with or without data of the user.

In \ Out											
	C	T	S	S	P	P	P	P	T	C	Recall
Circle	12	5	35	0	0	0	1	7	4	0	0.188
Triangle	3	47	10	4	0	0	0	0	0	0	0.734
Square	5	16	41	1	0	0	0	0	0	0	0.651
Star	0	11	21	25	0	5	0	0	0	0	0.403
Punch	0	1	0	1	14	18	4	21	0	0	0.237
Push door	0	1	0	1	14	18	4	21	0	0	0.305
Pull drawer	0	0	0	0	9	1	20	33	0	0	0.317
Pull door	0	0	0	0	9	1	20	33	0	0	0.524
Throw	0	5	0	0	2	8	0	17	18	17	0.269
Chop	0	5	0	0	2	8	0	17	18	17	0.254
Mean											0.388

(a) Text 1

In \ Out											
	C	T	S	S	P	P	P	P	T	C	Recall
Circle	53	3	8	0	0	0	0	0	0	0	0.828
Triangle	0	63	1	0	0	0	0	0	0	0	0.984
Square	0	19	46	0	0	0	0	0	0	0	0.708
Star	0	0	2	60	0	0	0	0	0	0	0.968
Punch	0	0	0	0	19	20	0	16	6	1	0.306
Push door	0	0	0	1	27	34	0	1	0	0	0.540
Pull drawer	0	0	0	0	0	0	61	0	0	0	1.000
Pull door	0	0	5	0	0	4	1	55	0	0	0.846
Throw	0	0	0	0	0	0	1	0	64	1	0.970
Chop	0	0	0	0	0	2	0	1	0	62	0.954
Mean											0.810

(b) Text 2

In \ Out											
	C	T	S	S	P	P	P	P	T	C	Recall
Circle	65	0	0	0	0	0	0	0	0	0	1.000
Triangle	0	65	0	0	0	0	0	0	0	0	1.000
Square	0	1	64	0	0	0	0	0	0	0	0.985
Star	0	0	5	59	0	0	0	1	0	0	0.908
Punch	0	0	0	0	56	11	0	0	0	0	0.836
Push door	0	0	0	0	1	56	2	6	0	0	0.862
Pull drawer	0	0	0	0	0	1	62	2	0	0	0.954
Pull door	0	0	0	0	0	3	2	60	0	0	0.923
Throw	0	0	0	0	1	1	0	0	62	1	0.954
Chop	0	0	0	0	0	0	0	0	0	65	1.000
Mean											0.942

(c) Text 3

In \ Out											
	C	T	S	S	P	P	P	P	T	C	Recall
Circle	65	0	0	0	0	0	0	0	0	0	1.000
Triangle	1	60	2	0	0	0	0	0	0	0	0.952
Square	0	5	59	0	0	0	0	0	0	0	0.922
Star	1	1	1	61	0	0	0	0	0	0	0.953
Mean											0.957

(d) Figure

In \ Out											
	C	T	S	S	P	P	P	P	T	C	Recall
Circle	63	0	0	0	0	0	0	0	0	0	1.000
Triangle	0	63	0	0	0	0	0	0	0	0	1.000
Square	0	0	64	0	0	0	0	0	0	0	1.000
Star	0	0	0	62	0	0	0	0	0	0	1.000
Punch	0	0	0	0	67	0	0	0	0	0	1.000
Push door	0	0	0	0	2	64	0	0	0	0	0.970
Pull drawer	0	0	0	0	0	0	66	0	0	0	1.000
Pull door	0	0	0	0	0	0	0	65	0	0	1.000
Throw	0	0	0	0	0	0	0	0	65	0	1.000
Chop	0	0	0	0	0	0	0	0	1	61	0.984
Mean											0.995

(e) Video

Table 3 Recognition results.

5. Conclusion

We have evaluated the effect of the gesture instructions on gesture recognition. In the evaluation, we captured data of ten kinds of gestures in three types of texts, figure, and video from 13 subjects, and examined change in movement for individuals, diversity among subjects, and recognition accuracy

by visual check and Dynamic Time Warping (DTW) algorithm. The evaluation results showed that the subjects who performed gestures correctly as the instructions got detailed, however specific words such as “door” and “ball” confused the subjects. In addition, the words “throw” and “chop” called multiple types of actions to the subjects’ mind since these actions have high degree of freedom.

As a future work, we plan to examine how end-users express gestures and how recognition system performs when the expressions are given back to them again.

Reference

- [1] Alexander, J., Han, T., Judd, W., Irani, P. and Subramanian, S. Putting your best foot forward: Investigating real-world mappings for foot-based gestures. In *Proceedings of CHI 2012* (2012), 1229–1238.
- [2] Bau, O. and Mackay, W.E. Octopocus: A dynamic guide for learning gesture-based command sets. In *Proceedings of UIST 2008*, ACM (2008), 37–46.
- [3] Bragdon, A., Zeleznik, R., Williamson, B., Miller, T. and LaViola, J. Gesturebar: Improving the approachability of gesture-based interfaces. In *Proceedings of CHI 2009*, ACM (2009), 2269–2278.
- [4] Fothergill, S., Mentis, H.M., Kohli, P. and Nowozin, S. Instructing people for training gestural interactive systems. In *Proceedings of CHI 2012* (2012), 1737–1746.
- [5] Graeme, S.C., Svetha V., Geoff A.W.W., and Hung H.B. Hierarchical recognition of intentional human gestures for sports video annotation. In *ICPR 2002*, vol. 2, IEEE (2002), 1082–1085.
- [6] HTK Hidden Markov Model Toolkit home page. <http://htk.eng.kcam.ac.uk/>.
- [7] Junker, H., Amft, O., Lukowicz, P., and Tröster G. Gesture spotting with body-worn inertial sensors to detect user activities. *The Journal of the Pattern Recognition Society* (2008), 2010–2024.
- [8] Keogh, E., Chu, S., Hart, D., and Pazzani, M. An online algorithm for segmenting time series. In *Proceedings of ICDM 2001*, IEEE (2001), 289–296.
- [9] Liu, J., Wang, Z., Zhong, L., Wiekramasuriya, J., and Vasudevan, V. uwave: Accelerometerbased personalized gesture recognition and its applications. In *Proceedings of PerCom 2009* (2009), 1–9.
- [10] Myers, C.S., and Rabiner, L.R. A comparative study of several dynamic time-warping algorithms for connected word recognition. In *The Bell System Technical Journal*, vol. 60(7) (1981), 1389–1409.
- [11] Rabiner, L.R. and Juang, B.H. . A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of IEEE*, IEEE (1989), 257–286.
- [12] Westeyn, T., Brashear, H., Atrash, A., and Starner, T. Georgia tech gesture toolkit: Supporting experiments in gesture recognition. In *Proceedings of ICMI 2003*, ACM (2003), 85–92.
- [13] Wireless Technologies, Inc. <http://www.wireless-t.jp/>.

(received Aug. 22, 2014)

Biography

Kazuya Murao (Member)



Kazuya Murao is an Assistant Professor at College of Information Science and Engineering, Ritsumeikan University, Japan. He received B.Eng, M.Info.Sci., and Ph.D. degrees from Osaka University in 2006, 2008, and 2010, respectively. Prof. Murao is working on wearable computing, ubiquitous computing, and context aware systems. He is a member of IEEE and IPSJ.

Tsutomu Terada (Member)



Tsutomu Terada is an Associate Professor at Graduate School of Engineering, Kobe University, Japan. He is also a PRESTO researcher at Japan Science and Technology Agency, Japan. He received B.Eng., M.Eng., and Ph.D. degrees from Osaka University in 1997, 1999, and 2003, respectively. Prof. Terada is working on wearable computing, ubiquitous computing, and entertainment computing. He is a member of IEEE, IPSJ, and IEICE.

