Speed Dependencies of Human Gesture Recognition

Junpei Endo
The School of Knowledge Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan
Email: junpei.endou@jaist.ac.jp

Masashi Inoue
Graduate School of Science and Engineering
Yamagata University
3-16, 4 Jyonan, Yonezawa, Yamagata, Japan
Email: mi@yz.yamagata-u.ac.jp

Abstract— Gestures are important communication medium. However, their semantic ambiguities make it difficult for computer systems to accurately recognize and deliver them. The interpretation of gestural motions is not even consistent in human-to-human communications. In this paper, we compare gestures in order to establish a method to reveal the influence of gesture speed on semantic interpretation. We captured whole-body motions as movie files with a depth camera and converted them into biological motion movie files. The speed of these motion images was then systematically altered, and a perceptual experiment was conducted on **11** participants. Using the results from the experiment, we identified speed dependencies of human gesture recognition.

I. Introduction

Gestures are one of the most important forms of non-verbal communication in the field of ubiquitous human-computer interaction. As an accessible means of computer input, gesture recognition has been studied intensively [1] and gesture interfaces are actively being developed. Additional studies on the role of gestures in affective communication and their associations with emotion in interaction with information systems have also been conducted [2]. They are also used to make humanoid robots and software agents appear more realistic or natural to human [3]. For such purposes, designers must have an idea of how gestures are perceived by humans when interacting with robots or agents. In developing gesture recognition systems, it is often assumed that gestures sustain their semantics regardless of their speed and spatial locations. However, different individuals generate the same gestures at different speeds and trajectories. To account for this, techniques that absorb time expansion or contraction and extract patterns in trajectory of motions such as dynamic time warping and hidden Markov models were introduced. Such techniques are useful when gestures are generated with clear intention, but difficulties in using gestures in human-computer interaction arise in their semantic ambiguities. In fact the interpretation of non-symbolic gestures is not even consistent among human [4]. Although the effect of gesture speed in determining expressivity has been studied[5], the area of semantic perception requires further exploration. Experiments have been conducted on how the interpretation of biological motion of gait is affected by speed[6]; however, categorical changes of semantic interpretation have not yet been discussed. Therefore, in this paper, we experimentally compare speed dependencies among gestures to establish a method to reveal the influence of gesture speed in semantic interpretation. We captured whole-body motions with a depth camera and converted the captured movie files into biological motions. The speed of these motion images was then systematically altered. A perceptual experiment was conducted with 11 participants and we used the results to identify speed dependencies of semantic gesture recognition between gestures. The findings serve as a basis for the future large scale investigation on the semantics of gestures.

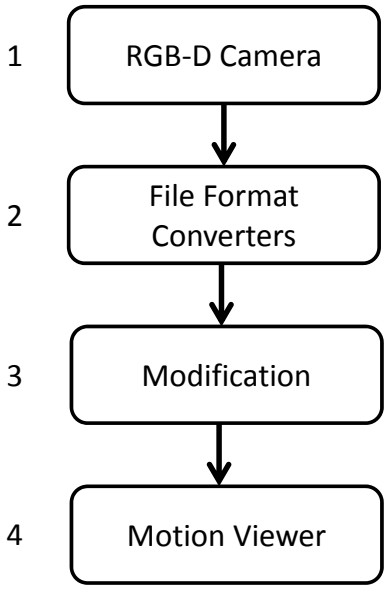

Fig. 1. Motion data preparation flow.

II. Methodology

A. Motion Capture

The stimuli were prepared using the steps shown in Figure 1. In the first step, a Kinect sensor was used to capture whole-body movements. Human motions were

recorded as XED files using the Kinect SDK toolkit [1]. In the second step, the file format was converted, and a modeling tool was used to import Kinect data and edit motions [2] [3]. In the third step, the speed of the captured motions was altered in accordance with the criteria outlined in Section II-C. The modified motion data in VMD format were then converted into BVH format with a motion creation software [4]. These BVH files were then configured by our own scripts. The details of configuration are outlined below and in Section II-C. Finally, the movie files were presented by a visualization software [5].

The motions shown to the experiment participants were presented in a form similar to biological motion [7]. As shown in Figure 2, human bodies were represented by point-light displays where the main joints were represented by small dots and were connected by lines to construct stick figures. The purpose of presenting wireframe figures instead of the original videos was to reduce the effects of other nonverbal information such as gender, body shape, and adornment on the perception of motions. Changes in motion speed may seem less unnatural in this stick figure representation. We made additional modifications to the size of figures and information of face direction. The figures were also enlarged to fill the screen for the visibility. Furthermore, the dot and line corresponding to nose positions were removed since their existence would give the viewer the impression that the figure was not facing forwards. The motions were repeated three times in each movie file, with a two-second interval between each repetition.

B. Gesture Selection

Two criteria were used to determine which gestural stimuli to include in the experiment. The first took into account the resolution and accuracy of Kinect sensors in capturing motions, since they are unable to detect gestures in small body parts such as fingers. Therefore, we decided to include only gestures that may be perceivable if presented as whole-body wireframe figures. The second criterion was that the gestures should contain some ambiguities. The reasoning behind this was if the selected gestures were recognized uniquely without any hesitation, they may not be affected by speed changes. Ideally, the participants should deliberate the meanings of the gestures rather than instantly interpret them. Iconic or deictic gestures were therefore avoided. To conduct a comprehensive experiment, we need to test larger collection of gestures. For the current preliminary experiment, we started with the following four gestures. They are selected on the basis of that they might be emotionally ambiguous. That is, their emotional impression differs in individuals. Although there are differences between semantic perceptions and emotional impressions, we arbitrarily selected gestures that might be semantically ambiguous. The target gesture selection should follow more systematic steps in the future.

A	Raising one's hand
B	Restraining something by hand
C	Kicking something on the floor
D	Being arrested and put handcuffs

They have different durations: gestures A takes 1.0 second, gesture B takes 2.5 seconds, gesture C takes 4.4 seconds, and gesture D takes 1.5 seconds. Gestures A, B, and D are interactive in that there is someone to whom the gesture is addressed or someone who stimulated the gesture. Despite the titles assigned to each gesture, the motions are not necessarily interpreted as such and there is room for different interpretations. In particular, we expected that the loss of hand information would make the interpretation of gestures in which hand shapes have a crucial role difficult.

C. Speed Variation

For each motion trajectory, we prepared two movie files with different movement speeds. The speeds of the four gestures was altered as follows: Gesture A was slowed down to one-fourth speed, gesture B was sped up by a factor of 2.5, gesture C was sped up by a factor of 2, and gesture D was slowed down to half speed. Our intention was to alter the speed of the gestures enough to be noticed by the participants, but not so much that they would appear unnatural. These speed rates were decided subjectively based on our observation for the purpose of the preliminary experiment. In future experiments, the speed should be systematically increased to cover the entire range. The speed changes were applied from the point when the hands or legs were at their home positions to when the gesture stroke ended. That is, in modified gestures, the original retraction speeds were left intact. We found that with the exception of gesture D, speed changes applied to retraction phases made the gestures unnatural. Speed changes were applied throughout each phase of gesture D, since this made the motion more natural than segmented speed control. The future challenge is to automatically determine the phases to be changed without losing naturality for different gestures when their speeds are altered .

[1] [Kinect Studio]:
http://msdn.microsoft.com/en-us/library/hh855389.aspx
[2] [MikuMikuDance]:
http://www.geocities.jp/higuchuu4/index_e.htm
[3] [MoggNUI]:
https://sites.google.com/site/moggproject/
[4] [LiveAnimation]:
http://www.drf.co.jp/liveanimation/index_en.html
[5] [BVHViewer]:
http://vipbase.net/bvhviewer/

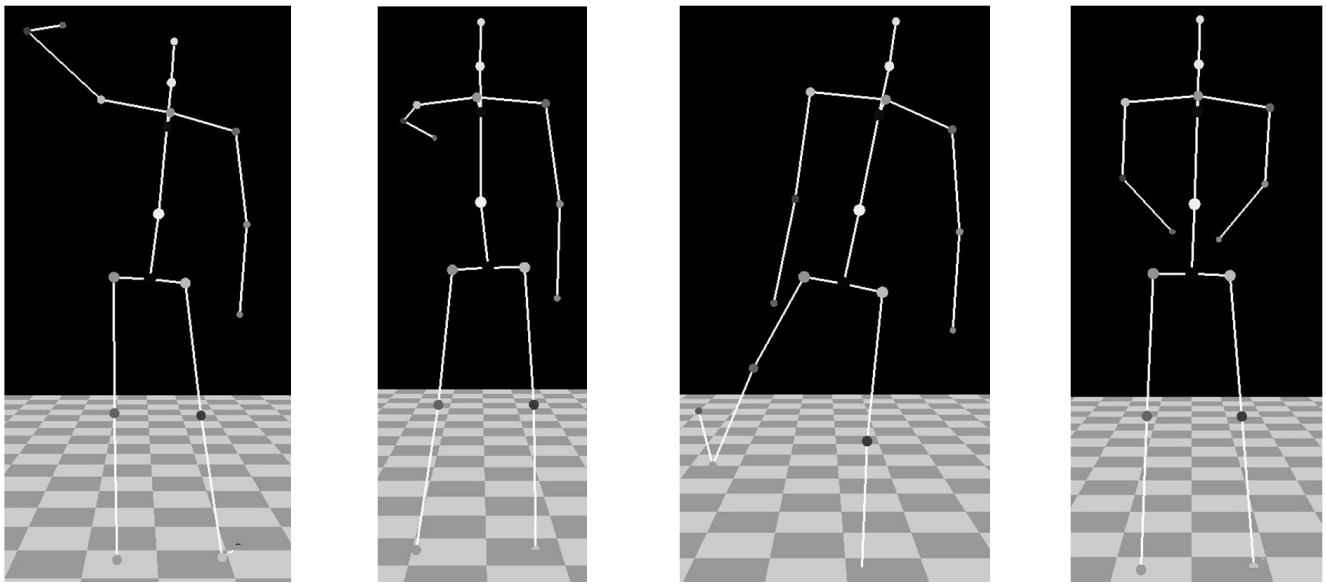

Fig. 2. Snapshots of the four motion movies presented to the participants.

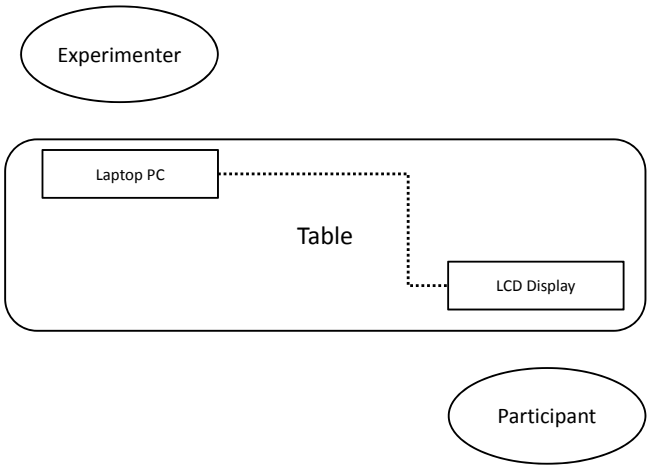

Fig. 3. Experiment room layout.

III. Experiment

A. Experimental Procedure

Eleven undergraduate students participated in the experiment. When participants entered the room, the experimenter verbally explained that they were going to be shown several gesture movie files. Since most participants were not familiar with the motion captured movies, we further explained how Kinect sensors captured human motions and motion data were then converted into movies. The participants were then asked to produce a stepping gesture in front of a Kinect sensor and the data were transformed into movies based on the same process used in the creation of the experimental stimuli. By doing so, the participants could understand the meaning of the movie being presented.

In the experiment, a participant was sequentially presented the motion stimuli and asked to fill in a questionnaire form. The experimenter controlled the computers and the participants watched the movie files displayed over an LCD. The layout of the experiment room is shown in Figure 3. All participants were presented each of four motions in either their original or in modified speed. For the same motion, gestures at different speeds were randomly presented to different participants in order to avoid order effect. In the questionnaire, participants wrote down no more than three interpretations of the presented motion in order of their confidence level. If a participant was not able to think of a meaning associated with a gesture, they could write "I have no idea." They were asked to fill in the form within a minute after the end of each motion presentation. When they finished writing, they could move to the next gesture prior to the end of one minute slot.

B. Experimental Results

The goal of the experiment was to identify the change in semantic recognition of gestures when the speeds of their movements are changed. The participants recorded their interpretations as free descriptions and this resulted in varied responses. Therefore, we categorized the provided responses before commencing with an quantitative analysis of the data. For example, interpretations of gesture A yielded four similar descriptions: "raising his hand," "raising hand," "raising his right hand," and "raising hand motion." These responses were categorized as the same interpretation – "Hand raising." Other descriptions such as, "taking something from the shelf," "calling someone," and "taking a backswing" were categorized as three different interpretations. Similarly, two descriptions for gesture D – "receiving an award certificate" and "holding

TABLE I

Two most popular interpretations given by participants for the four gestures.

Gesture	First Interpretation	Second Interpretation
A	Hand raising	Greeting
B	Elbow strike	Pulling
C	Kicking	Throwing shoes
D	Receiving something	Kneading waist

TABLE II

Interpretation frequencies in the two most popular categories for each of the four gestures.

Gesture	Original Speed		Modified Speed	
	First	Second	First	Second
A	2	3	4	2
B	1	2	1	1
C	6	1	5	1
D	3	0	0	4

something with both hands" – were both grouped into a "Receiving something" category. Also, "kneading his waist," "kneading his back," and "moving his hands to waist" were categorized into "Kneading waist." Among the re-organized categories, we chose the two most popular ones from each gesture for the analysis, which are listed in Table I. We then tallied the participants' interpretations and summarized them in Table II. The total numbers are different for each gesture because only descriptions from the two most popular categories were included.

From the data in Table II, we first determined which gestures were ambiguous. Ambiguous gestures received recognition numbers evenly partitioned among typical interpretations. In the case of gesture B, the motion portrayed could reasonably be seen as elbow strike or pulling. Such ambiguous gestures in which the difference between the counts for two interpretations is smaller than 2 are marked with = sign in Table III. Similarly, if one interpretation was clearly more popular than the other, the relationships are marked by either $>$ or $<$ signs. $>$ means the first interpretation at the original speed is preferred more than the second interpretation. $<$ means the second interpretation at the original speed is preferred more than the first interpretation. There are two types of speed dependencies that can influence the recognition of a gesture. The first is that the variation in recognitions moves towards either uniformity (same semantic interpretation among participants) or diversity (different interpretation among participants). Gesture A is an example of this change; participants' interpretations became more uniform when the speed of gesture slowed down as shown in Table III. The second dependency is when the $<$ and $>$ signs are inverted. In this case, the change of speed alters the semantics of gestures. An example of this can be seen in Table III, where gesture D matches this criteria. At original speed, the participants interpreted gesture D as

TABLE III

Magnitude relationship of interpretation frequencies between gestures at original speed and gestures at modified speed. = signs mean two interpretations were equally preferred, $>$ and $<$ signs indicate one of two interpretations was preferred.

Gesture	Original speed	Modified speed
A	=	$>$
B	=	=
C	$>$	$>$
D	$>$	$<$

the gesture for receiving something, but when the motion speed was reduced, it was perceived as the gesture for massaging oneself.

IV. Conclusion

This paper described our method for identifying gestures that are interpreted differently at different speeds. With deeper understanding of human gestures, naturally occurring gestures can be used as input to computers rather than pre-defined motion sets as gesture commands. Current human-computer interaction systems that utilize user gestures as gesture commands will be benefited. For example, button push gestures can be used as the message for pushing buttons while conventional systems require arm swing motion for the same operation. We used a Kinect sensor and associated tools to motion capture gestures and transformed the original videos into motion movie files featuring wireframe bodies. From these wireframe movies, subsequent movies featuring the same gesture and increased and decreased speeds were created. Both the original and modified motions were presented to experiment participants and we asked them to interpret each gesture. As the result of experiment, we found two types of speed dependencies in gestures in terms of semantic interpretation. One type of gestures were interpreted similarly at certain speed but become ambiguous when their speed changes. Another type of gestures were interpreted differently based on their speeds. The experiment in this paper is a preliminary one to test the idea of using modifiable motion images in measuring speed dependencies of gesture recognition. Our method can be applied to a larger scale experiment that involves a higher number of gesture categories and systematic speed varieties in order to identify more gestures whose recognitions are dependent on speed.

An important factor that we have not tested is the influence of body orientation on the recognition of gestures. In our experiment, we only included videos featuring front-facing figures. Figures standing sideways, for instance, may elicit varied interpretations. Another factor that was not tested was the interaction between verbal and non-verbal information. It is known that the comprehension of gestures can be helped by speech [8]. When gestures are presented with words or phrases, it is likely to be interpreted within a narrower scope. The third factor we

did not heavily take into consideration was the degree of confidence in the participants' responses. Although we collected subjective confidence scores from the participants, we did not use them in the current analysis since the number of data was considered too small for the statistical analyses. The integration of continuous values into the modeling of human gesture recognition could make our knowledge more usable in designing ubiquitous human-computer interaction systems. It is argued that there is a difference in the perception of gestures when they are produced by humans and by robots[9]. We believe that the motions presented in our experiment were recognized as the gestures generated by human and it would be interesting to examine the perceptions garnered from applying our gestures to non-human agents.

Acknowledgment

This work was done while the first author was at Yamagata University. This research was partially supported by the Grant-in-Aid for Scientific Research 24500321.

References

[1] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 677–695, July 1997.

[2] D. M. Bustos, G. L. Chua, R. T. Cruz, J. M. Santos, and M. T. Suarez, "Gesture-based affect modeling for intelligent tutoring systems," in Proceedings of the 15th international conference on Artificial intelligence in education, 2011, pp. 426–428.

[3] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wöllmer, "Building autonomous sensitive artificial listeners," IEEE Transactions on Affective Computing, vol. 3, no. 2, pp. 165–183, Apr. 2012.

[4] M. Mahmoud and P. Robinson, "Interpreting hand-over-face gestures," in Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II, 2011, pp. 248–255.

[5] B. Hartmann, M. Mancini, and C. Pelachaud, "Implementing expressive gesture synthesis for embodied conversational agents," in Proceedings of the 6th international conference on Gesture in Human-Computer Interaction and Simulation, 2006, pp. 188–199.

[6] J. A. Beintema, A. Oleksiak, and R. J. A. van Wezel, "The influence of biological motion perception on structure-from-motion interpretations at different speeds," Journal of Vision, vol. 6, no. 7, pp. 712–716, June 2006.

[7] G. Johansson, "Visual perception of biological motion and a model for its analysis," Perception & Psychophysics, vol. 14, no. 2, pp. 201–211, June 1973.

[8] S. D. Kelly, A. Özyürek, and E. Maris, "Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension," Psychological Science, vol. 21, no. 2, pp. 260–267, 2010.

[9] C. J. Hayes, C. R. Crowell, and L. D. Riek, "Automatic processing of irrelevant co-speech gestures with human but not robot actors," in Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction, 2013, pp. 333–340.