Method of Determining Training Data for Gesture Recognition Considering Decay in Gesture Movements

Gaku Yoshida, Kazuya Murao, Tsutomu Terada, Masahiko Tsukamoto
Graduate School of Engineering, Kobe University
Kobe, Hyogo, Japan
{*g-yoshida@stu., murao@eedept., tsutomu@eedept., tuka@*}*kobe-u.ac.jp*

***Abstract*—Mobile phones and video game controllers using gesture recognition technologies enable easy and intuitive operations, such as those in drawing objects. Gesture recognition systems generally require several samples of training data before recognition takes place. However, recognition accuracy deteriorates as time passes since the trajectory of the gestures changes due to fatigue or forgetfulness. We investigated the change in gestures and fast found that several samples of gestures were not suitable for training data. Therefore, we propose two methods of finding appropriate data for training. We confirmed that the proposed methods found better training data than the conventional method from the viewpoints of the number of data collected and the accuracy of recognition.**

***Keywords*-gesture recognition; training data selection; accelerometer;**

I. Introduction

Downsizing of computers has led to mobile and wearable computing that has recently attracted a great deal of attention. In particular, accelerometers are installed in most current devices, such as iPhones and Android-powered devices, and the video game controllers for Wii or PS3, which enable easy and intuitive operations. In addition, life logs can be enriched by recording gesture movements in daily life with wearable sensors.

Gesture recognition systems generally have to be trained with users' gesture data before use, and they recognize unknown gestures by comparing them with the training data. The simplest and most common way of training systems is to use a few samples of gesture data before using the systems for the first time. However, recognition accuracy could deteriorate since the gestures to be recognized change due to user conditions, such as him/her forgetting the original gestures and day-to-day fatigue. As far as we know, no effective methods of training systems that have taken change into consideration in future gesture motions have yet been reported.

We investigated the change in gesture motions by repeating specific gestures 200 times a day for seven days. As gesture motions were changed by repetition in the experiment, the first several samples were not suitable for training data such as those in the conventional method. Therefore, we propose two methods of finding training data taking changes in future gesture motions into account. We confirmed that the new methods found training data that were robust to temporal changes in the early stages of training.

This paper is organized as follows. Section 2 introduces related work. Section 3 describes a preliminary experiment that we conducted to evaluate what effect changes would have on gesture motions. Section 4 presents our proposed approaches. The experiment we conducted to evaluate the accuracy of our methods are described and the results are presented and discussed in Section 5. The key points are summarized and future work is mentioned in Section 6.

II. Related work

Many studies on gesture recognition using accelerometers have been reported. Murao et al. [1] evaluated recognition accuracy for 27 kinds of gestures with nine accelerometers and nine gyroscopes on a board and demonstrated the differences in recognition accuracy by changing the number, positions, and kinds of sensors and the number and kinds of gestures. Agrawl et al. [2] proposed a system that recognized the alphabet written in the air with a cell phone. Acceleration data were converted to spatial motion. They achieved 83% recognition accuracy by adhering to some restrictions. The method proposed by Chambers et al. [3] was used to annotate video-recorded activities by gesture recognition using an accelerometer mounted on the wrist since it is difficult to annotate video only by analyzing it. They conducted three types of Kung-fu gestures, i.e., cutting, punching, and elbowing, resulting in one mistake in 30 trials using hidden Markov models (HMMs). The system proposed by Junker et al. [4] recognized ten daily short actions, such as pushing a button and drinking, and achieved approximately 80% precision and recall. The training data in these studies were collected before recognition without taking into consider, on changes in gesture motions caused by user conditions.

In contrast to these studies, Liu et al. [5] focused on changes in daily gesture motions. They recognized eight kinds of gestures including those in drawing a line or a circle (recommended by Nokia Research Institute), with a 3-axis accelerometer. They captured more than 4,000 samples for eight subjects over a long period, using dynamic time-warping (DTW) [6] as a recognition algorithm. An accuracy

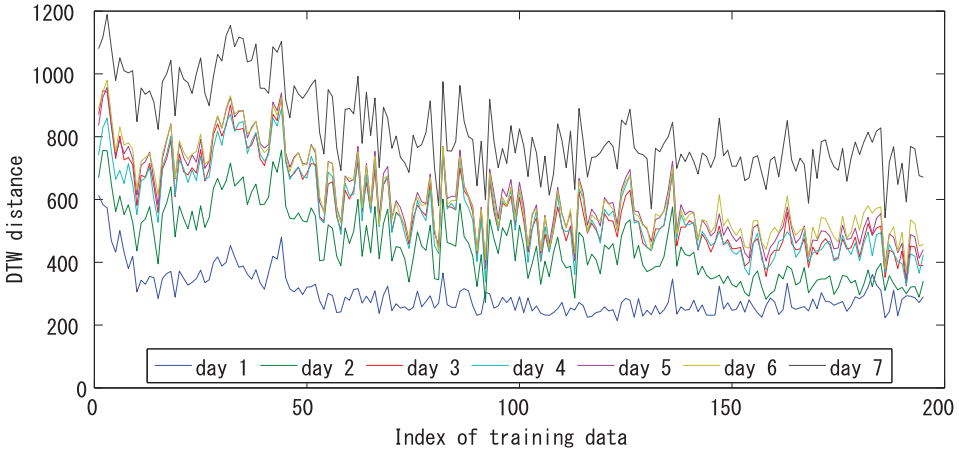

Figure 1. Average DTW distance for *throwing a ball* gesture for subject 1.

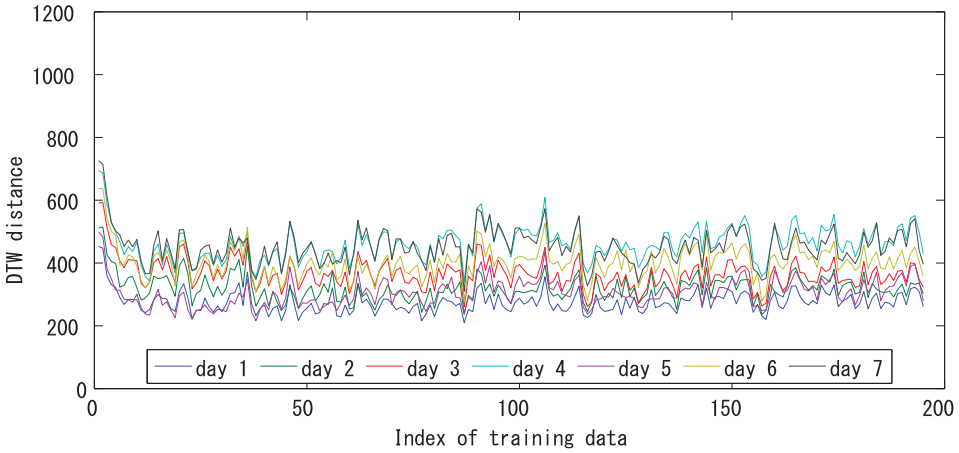

Figure 2. Average DTW distance for *drawing a star* gesture for subject 1.

of 98.6% was achieved by successively renewing the training data. However, a ground truth was required when renewing the training data, which did not fundamentally solve the problem. In addition, it was not clear whether the system was effective against changes in gesture motions throughout the day since training data were renewed once a day.

III. Preliminary Experiment

We conducted a preliminary experiment to investigate the change in gesture motions over a long period.

A. Setup and procedure

Data on two kinds of gestures: *throwing a ball* and *drawing a star in the air*, were captured continuously 200 times in 3-second intervals for seven days for five male subjects 22–23 years old who had a 3-axis accelerometer attached to their wrist. The reason for this setup was that gesture motions would change as they were repeated many times and the day progressed. The change in gesture motions was investigated with the following procedure. First, 200 sequences for seven days for each gesture $x_{i,j}$ were collected, where i is an index of the sequences on a day $(1 \leq i \leq 200)$ and j is a day $(1 \leq j \leq 7)$. Then, the distances between training data $x_{i,1}$ and testing data $x_{i',j'}$ were calculated with a dynamic time-warping (DTW) algorithm for $1 \leq i' \leq 200$ and averaged, iterating this calculation for $1 \leq i \leq 200$ and for $1 \leq j' \leq 7$.

The accelerometer used was a WAA-006 sensor, made by Wireless Technologies Inc. [7]. The sampling frequency was 50 Hz.

B. Results

The results for the *throwing a ball* and *drawing a star in the air* gestures are plotted in Figures 1 and 2. The vertical axis indicates the average DTW distance and the horizontal axis indicates an index of the sequence for the training data. For example, the line "day 2" in Figure 1 indicates the average DTW distances between the 200-sample data on the second day and all data on the first day $x_{i,2}(1 \leq i \leq 200)$ and i corresponds to the horizontal axis for the *throwing a ball* gesture. The value on the line at the point where the horizontal axis is 100 denotes the average DTW distances

Table I
RECOGNITION RESULTS FOR *throwing a ball* GESTURE.

Day	Throw	Star	Circle	Triangle	Square	Punch	Chop	Slap	Vertical line	Horizontal line	Accuracy [%]
1	1000	0	0	0	0	0	0	0	0	0	100
2	1000	0	0	0	0	0	0	0	0	0	100
3	1000	0	0	0	0	0	0	0	0	0	100
4	997	0	0	0	0	3	0	0	0	0	99.7
5	1000	0	0	0	0	0	0	0	0	0	100
6	999	0	0	0	0	1	0	0	0	0	99.9
7	1000	0	0	0	0	0	0	0	0	0	100

Table II
RECOGNITION RESULTS FOR *drawing a star* GESTURE.

Day	Throw	Star	Circle	Triangle	Square	Punch	Chop	Slap	Vertical line	Horizontal line	Accuracy [%]
1	0	998	1	1	0	0	0	0	0	0	99.8
2	0	867	0	133	0	0	0	0	0	0	86.7
3	0	843	0	157	0	0	0	0	0	0	84.3
4	0	853	0	147	0	0	0	0	0	0	85.3
5	4	918	0	78	0	0	0	0	0	0	91.8
6	0	817	0	183	0	0	0	0	0	0	81.7
7	0	789	0	197	0	1	0	0	13	0	78.9

between the 200-sample data on the second day and 100th data on the first day. The distance between the same data was neglected in the calculations for day 1. The large value for the DTW distance indicates that the data at the point are different to the 200-sample data and that these are not suitable for training data. Moreover, the large fluctuations on the line indicates that the data around the point are not stable and that their quality differs vastly according to the point.

C. Discussion

The forms of the charts were categorized into two types from the results: unvaried lines and decreasing lines. There is an example of a decreasing line in Figure 1 and of an unvaried lines in Figure 2. The DTW distances at the beginning of both lines are greater than those in the latter half, indicating that gestures had changed through the experiment and that distances were greater when using the first few samples as training data, which is the conventional method. Eventually, the distance decreased in relation to the other gestures, resulting in misrecognition.

Moreover, the DTW distance converged as the number of trials increased, as shown in Figure 1. This is because the form of the subject's gestures stabilized since the subject had become accustomed to the gestures or gesture forms that varied with those that had less fatigue. Most of the results for the *drawing a star in the air* gesture were unvaried lines. This is because the subjects usually did not perform the gesture; therefore, the forms did not stabilized even with repetition or because *drawing a star* gesture involved a long motion; therefore, the form barely stabilized due to fatigue.

The fluctuations in the lines in Figures 1 and 2 decreased as the index of training data increased, indicating the gesture motions were stabilizing. We assumed from the results that the distance converged and its fluctuations decreased at the best point of the data for the training. Extremely accurate training data could be obtained by stopping data collection on a point where the line of the distance converged and stabilized for decreasing lines and by stopping data collection in the early phases for unvaried lines.

We calculated the recognition accuracy to conduct further investigations into the effect of change on gesture motions. The subjects performed eight kinds of gestures: *drawing a circle*, *drawing a triangle*, *drawing a square*, *punching*, *chopping*, *slapping*, *drawing a horizontal line*, and *drawing a vertical line* ten times each. These data and the data for the *throwing a ball* and *drawing a star* gestures for the first ten sequences of 200 sequences on the first day were used for training. Recognition took place for the 200 sequences for the *throwing a ball* and *drawing a star* gestures as follows. The DTW distances between testing data to be recognized and all the training data were calculated and gesture labels were annotated with data whose DTW distance was the shortest overall. Tables I and II summarize the recognition results for the *throwing a ball* and *drawing a star* gestures for 1,000 samples (200 samples $\times$ 5 subjects) per day. The rows indicate the number of outputs for the gestures.

The recognition accuracy for the *throwing a ball* gesture was over 99.7%. The recognition accuracy for the *drawing a star* gesture on the first day was 99.8%, while the recognition accuracy dropped to 78.9% on the seventh day. These results indicate that retaining the first few samples for training causes misrecognition as does the conventional method, which deteriorates interface usability or system accuracy using gesture recognition technology.

IV. Proposed method

The results from the preliminary experiment clarified that the point where the average distance and its variance was least was most appropriate for training data. However, it was not possible to predict the point by using uncollected future data. We propose two methods of finding the appropriate point when data are collected for training in real time.

A. Proposed method 1

We confirmed that the average DTW distance and its variance were least at the most appropriate point for training data. First, this method was used to collect gesture data n times, then to calculate the average μ_n and variance σ_n of DTW distances between data x_{n-10} and data x_{n-i} for $0 \leq i \leq 9$, where x_i is an ith gesture sequence.

$$\mu_n = \frac{\sum_{i=0}^{9} DTW(x_{n-i}, x_n)}{10}$$

and

$$\sigma_n = \frac{\sqrt{\sum_{i=0}^{9} \{DTW(x_{n-i}, x_n) - \mu_n\}^2}}{10},$$

where $DTW(x, y)$ is a function that calculates the DTW distance between sequences x for testing and y for training.

The system determines that the gesture motion has converged and stops collecting training data when both μ_n and σ_n satisfy four conditions.

- $\mu_n < \mu_0$,
- $\sigma_n < \sigma_0$,
- $\mu_{n-i} < \mu_n \times (1 + \alpha)$ for $i = 1, \cdots, 9, and$
- $\sigma_{n-i} < \sigma_n \times (1 + \alpha)$ for $i = 1, \cdots, 9$.

We set $\alpha = 0.01$ and the proposed method requires at least 20 gesture samples since 10 samples are used for calculating the average and variance and 10 samples are used for accessing the four conditions. These values were determined from our pilot study.

B. Proposed method 2

We confirmed that gesture motions were diverse in the first few samples and were going to remain unvaried as gestures were repeated from the preliminary experiment. We assumed that gesture motions were meant to be stable and outlying motions barely appeared when the distances between the first few samples and the current sample were continuously close. The algorithm is as follows. The DTW distance of x_{n-i} for $0 \leq i \leq 4$ is calculated with x_i for $1 \leq i \leq 5$ for training, and average μ'_n is calculated.

$$\mu'_n = \frac{\sum_{i=0}^{4} DTW(x_{n-i}, Y = \{x_j | \in 1 \leq j \leq 5\})}{5},$$

where $DTW(x, Y)$ is a function that calculates the DTW distance between sequences x for testing and a set of sequence Y for training.

The system stops collecting data for training when μ'_n meets the following condition.

- $|\mu'_{n-i} - \mu'_n| < \mu'_n \times \beta$ for $i = 1, \cdots, 5$

We set $\beta = 0.1$ and the proposed method requires at least 10 gesture samples since five samples are used for calculating the average and five samples are used for determining the condition above. These values were determined from our pilot study.

V. Evaluation

We calculated the DTW distance between the data for *throwing a ball* and *drawing a star* gestures captured in the preliminary experiment and the training data obtained with proposed methods 1 and 2 and three other methods of comparison to evaluate the effectiveness of our proposed methods: Comparison method 1 uses the first five samples on the first day for training, which is the conventional method. Comparison method 2 uses the data for training at the point where the DTW distance becomes less than 110% of the last value on the first day, which is the baseline method. Comparison method 3 uses the data for the first five samples and updates the training data every day, which would perform well but forces the user into harder tasks. Tables III and IV list the average DTW distance for *throwing a ball* and *drawing a star* gestures for the proposed methods and the comparison methods. The parenthetic values indicate the number of training data obtained until each method stopped collecting data.

The DTW distance for the proposed method 1 was less than that for comparison method 1, but was greater than that for comparison methods 2 and 3 in some cases. This is because proposed method 1 stopped collecting data at the point where the data were continuously stable but had not converged. Moreover, the DTW distance for comparison method 2 was greater than that for comparison method 3, which means that updating the training data every day improves accuracy.

Comparing the results for subjects 1 and 2 for the *throwing a ball* gesture, the DTW distance for proposed method 1 was greater than that for comparison methods 2 and 3, while fewer samples were collected, than those for comparison method 1. The results for subjects 3 and 4 for proposed method 1 were comparable to those for comparison method 2 and less than those for the comparison method 3 in some cases. The results for subject 5 for proposed method 1 were greater than those for comparison method 2 but less than those for comparison method 3.

The results for proposed method 1 for subjects 2 and 5 for the *drawing a star* gesture were greater than those for the comparison methods 2 and 3 but less than those for the comparison method 1 with the small number of training data collected. The results for subjects 1, 3, and 4 for proposed method 1 were comparable to those for the comparison method 2 and less than those for comparison method 3 in

Table III
AVERAGE DTW DISTANCE AND NUMBER OF DATA COLLECTED FOR *throwing a ball* GESTURE OVER SUBJECTS.

Subject		Pro 1	Pro 2	Com 1	Com 2	Com 3
	# of data	(48)	(20)	(5)	(158)	(5*)
	Day 2	395	395	560	255	372
	Day 3	557	557	730	349	435
1	Day 4	525	525	662	329	331
	Day 5	579	579	738	373	309
	Day 6	608	608	747	404	287
	Day 7	817	817	974	562	373
	# of data	(41)	(15)	(5)	(112)	(5*)
	Day 2	750	790	839	548	709
	Day 3	903	916	997	724	592
2	Day 4	741	768	821	612	759
	Day 5	754	768	834	604	606
	Day 6	746	748	832	618	629
	Day 7	760	760	893	631	624
	# of data	(123)	(165)	(5)	(68)	(5*)
	Day 2	285	282	387	299	393
	Day 3	331	324	444	356	355
3	Day 4	335	330	448	359	344
	Day 5	362	358	431	393	352
	Day 6	367	357	480	403	307
	Day 7	400	391	482	425	376
	# of data	(96)	(21)	(5)	(87)	(5*)
	Day 2	427	500	658	427	560
	Day 3	474	541	699	474	419
4	Day 4	439	495	628	440	442
	Day 5	475	531	611	475	480
	Day 6	452	531	676	452	400
	Day 7	428	491	603	428	431
	# of data	(58)	(13)	(5)	(122)	(5*)
	Day 2	520	587	615	456	566
	Day 3	496	584	618	432	506
5	Day 4	538	595	673	485	747
	Day 5	524	646	695	465	586
	Day 6	564	655	702	496	607
	Day 7	569	706	736	484	601

* First five samples per day.

Table IV
AVERAGE DTW DISTANCE AND NUMBER OF DATA COLLECTED FOR *drawing a star* GESTURE OVER SUBJECTS.

Subject		Pro 1	Pro 2	Com 1	Com 2	Com 3
	# of data	(22)	(30)	(5)	(48)	(5*)
	Day 2	271	257	389	231	464
	Day 3	316	308	444	289	356
1	Day 4	337	336	497	331	336
	Day 5	217	210	313	205	293
	Day 6	315	311	485	296	284
	Day 7	357	355	501	340	291
	# of data	(26)	(26)	(5)	(48)	(5*)
	Day 2	460	460	524	423	363
	Day 3	546	546	627	506	405
2	Day 4	537	537	503	497	503
	Day 5	511	511	558	475	439
	Day 6	517	517	559	503	541
	Day 7	494	494	560	464	467
	# of data	(53)	(48)	(5)	(69)	(5*)
	Day 2	285	282	387	299	393
	Day 3	362	366	524	329	400
3	Day 4	398	402	513	348	311
	Day 5	408	416	500	357	492
	Day 6	426	438	603	368	412
	Day 7	501	525	662	443	353
	# of data	(55)	(21)	(5)	(60)	(5*)
	Day 2	375	407	518	373	608
	Day 3	399	493	612	398	412
4	Day 4	371	479	649	371	333
	Day 5	369	471	664	369	341
	Day 6	437	527	692	437	354
	Day 7	531	678	893	531	347
	# of data	(96)	(15)	(5)	(103)	(5*)
	Day 2	386	679	681	374	409
	Day 3	508	940	940	477	392
5	Day 4	501	942	942	479	344
	Day 5	428	784	784	407	363
	Day 6	433	824	824	403	358
	Day 7	450	830	830	433	367

* First five samples per day.

some cases. The results demonstrated proposed method 1 was most effective in many cases.

The results for proposed method 2 were greater than those for comparison method 2 in most cases but less than those for comparison method 1. The distance for proposed method 2 was greater than that for proposed method 1, while the number of training data collected was smaller. This is because similar forms of the gesture continuously appeared before real convergence; therefore, the system failed to collect data that appeared in subsequent processes. We were able to reduce the number of training data collected compared with proposed method 1, but we need to consider cases where gestures change with two or more convergence points.

The distance for proposed method 2 for the *throwing a ball* gesture was greater than that for comparison method 2 for subjects 1, 2, 4, and 5. Compared to comparison method 1, both the distance was less and the number of data collected was smaller. Although numerous training data were collected for subject 3, the DTW distance was less than that for comparison method 2 and less than that for comparison method 3 in some cases.

The DTW distance for the *drawing a star* gesture for subject 5 for proposed method 2 was comparable to that of comparison method 1 and greater than that for comparison methods 2 and 3. Therefore, proposed method 2 was not effective for subject 5. The results for proposed method 2 were greater than those for comparison method 2 for subjects 1, 2, 3 and 4, but less than those for the comparison method 3 in some cases. Therefore, proposed method 2 was more effective.

Summarizing the results, both proposed methods 1 and 2 performed well. However, there was a trade-off between the quality of training data and the time to finish collecting them. The quality of training data for proposed method 1 was better than that for proposed method 2, while the data for the proposed method 2 were collected more rapidly than those for proposed method 1; therefore, these methods were selectively used depending on applications.

VI. Conclusion

We proposed methods of finding appropriate points for training data by calculating the DTW distance between collected data in real time. Compared with the conventional method, our method found better training data, with which the DTW distance was less than that for data arriving in the future. However, we have to improve the performance of the proposed methods because their DTW distances were greater than that for the method that uses data at ideal convergence points and that updates training data every day.

We plan to evaluate other kinds of gestures and conduct further investigations into the effect of fatigue and forgetfulness in future work.

Acknowledgment

This research was supported in part by a Grant in aid for Precursory Research for Embryonic Science and Technology (PRESTO) from the Japan Science and Technology Agency and by a Grant-in-Aid for Scientific Research (A)(20240009) and Scientific Research for Young Scientists (B)(24700066) from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

[1] K. Murao, T. Terada, A. Yano, and R. Matsukura: Evaluating Gesture Recognition by Multiple-Sensor-Containing Mobile Devices, *Proc. of International Symposium on Wearable Computers (ISWC 2011)*, pp. 55–58 (Oct. 2011).

[2] S. Agrawal, I. Constandache, S. Gaonkar, R. Choudhury, K. Caves, and F. DeRuyter: Using Mobile Phones to Write in Air, *Proc. of The International Conference on Mobile Systems, Applications, and Services (Mobisys 2011)*, pp. 15–28 (June/July 2011).

[3] G. S. Chambers, S. Venkatesh, G. A. W. West, and H. H. Bui: Hierarchical Recognition of Intentional Human Gestures for Sports Video Annotation, *Proc. of International Conference on Pattern Recognition (ICPR 2002)*, pp. 1082–1085 (Aug. 2002).

[4] H. Junker, O. Amft, P. Lukowicz, and G. Tröster: Gesture Spotting with Body Worn Inertial Sensors to Detect User Activities, *Pattern Recognition*, pp. 2010–2024 (2008).

[5] J. Liu, Z. Wang, L. Zhong, J. Wiekramasuriya, and V. Vasudevan: uwave: Accelerometer Based Personalized Gesture Recognition and Its Applications, *The IEEE Pervasive Computing and Communication (PerCom 2009)*, pp. 1–9 (June 2009).

[6] C. Myers and L. R. Rabiner: A Comparative Study of Several Dynamic Time-warping Algorithms for Connected Word Recognition, *The Bell System Technical Journal*, Vol. 60, pp. 1389–1409 (1981).

[7] Wireless Technologies, Inc.: http://www.wireless-t.jp/.