Pointing Gesture Recognition using Compressed Sensing for Training Data Reduction

Masahiro Iwasaki
Tokyo University of Agriculture and Technology
2-24-16 Naka-cho
Koganei, Tokyo Japan
masahirocky@gmail.com

Kaori Fujinami
Tokyo University of Agriculture and Technology
2-24-16 Naka-cho
Koganei, Tokyo Japan
fujinami@cc.tuat.ac.jp

UbiComp'13 Adjunct, September 8–12, 2013, Zurich, Switzerland.

http://dx.doi.org/10.1145/2494091.2495985

Abstract
In this paper, we investigate training data reduction for the pointing gesture recognition with compressed sensing. The pointing gesture is one of activities during pointing and calling that is carried out by workers to keep occupational safety and correctness. Compressed sensing is used for gesture recognition and considered the impacts of the gesture duration difference among user. However, the different force among users may affect to the recognition. As a result of the experiment, F-measure is improved 0.18 compared with the DTW even only the data obtained from others is used. Moreover, we found that the user-dependency varies for each subject. Therefore, we tested to recognize the pointing gestures of all subjects by using the training data of only specific users. The test showed that the recognition model with training data from 4 specific subjects provided the same accuracy as the one from 11 subjects. This result suggested the feasibility of reduction for subjects who need to acquire the training data.

Author Keywords
Ubiquitous Computing; Gesture Recognition;

ACM Classification Keywords
C.3 [Special-Purpose and Application-Based Systems]: Signal processing systems

Introduction

Pointing and calling is activity invented in Japan, and used to avoid the human errors at workplaces[9]. For a facilitated periodic assessment of pointing and calling, we proposed a system to recognize the pointing and calling. To develop the system, four targets are selected. These target movements are required to achieve the positive effect of the pointing and calling. They are pointing gesture (arm swing), gaze and pointing direction, vocalization, and location. Then pointing gesture is divided into two parts. Figure 1 shows the order of the pointing and calling activity defined by JICOSH[6]. During pointing and calling, the arm points a target twice, i.e. (2) and (4). Therefore, we devided the sequence into two parts between (2) and (3). The former part is named as Pointing gesture A, while the latter part is Pointing gesture B.

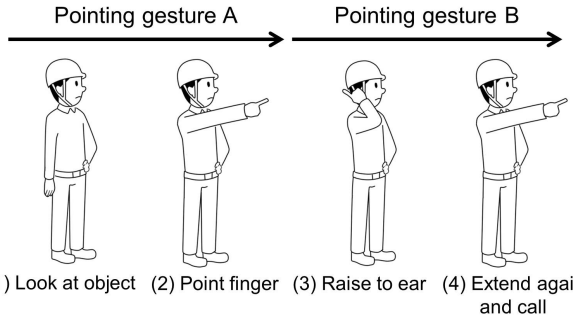

Figure 1: Pointing and calling activity sequence.

The Pointing gesture A and B are considered as gesture behaviors because these are movements of the arm. In the literature of gesture recognition, inertial sensors such as 3-axis accelerometer and/or 3-axis gyroscope are taking advantage. By using the time series data obtained from inertial sensors, Dynamic Time Warping (DTW)[12][8] and hidden Markov models (HMM)[14][10][11][7] are generally utilized. In this project, we attached an inertial sensor to wrist for acquiring the data. Also, we used DTW method because the large number of training data is required in order to determine all models in the HMM[9]. However, Gillian, et al.[8] who applied DTW in multivariate temporal musical gestures stated the need for data acquisition of more than 11 times for each gesture in the recognition of an accuracy over 90%. Therefore, the training data are required from original user who use the recognition system, which makes burden on users before the use of the system.

Motivation

For pointing and calling recognition system described in the previous section, the challenge is to investigate user-independent pointing gesture recognition. User-independent gesture recognition system is feasible to recognize gestures of specific user by using training data of another person. The challenge realizes the recognition without training data acquisition for each user before use of the system. Without data acquisition, the burden on the user decrease compared to DTW and HMM. However, recognition systems are basically dependent on the user that need template data generated by training data from the original user[12][8].

Figure 2 shows the duration of Pointing gesture A and B by 12 subjects for 20 times repetition. There are variations in the required time within subjects in addition to among subjects. The duration depends on the user and the speed of the movement. Accordingly, the time sequenced data can be either compressed or stretched.

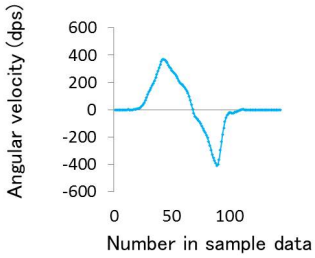

Figure 3: Example of time series data for gesture.

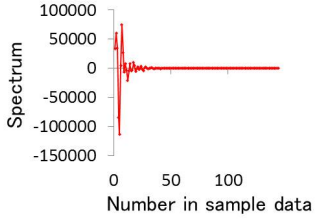

Figure 4: DCT coefficients of the gesture data.

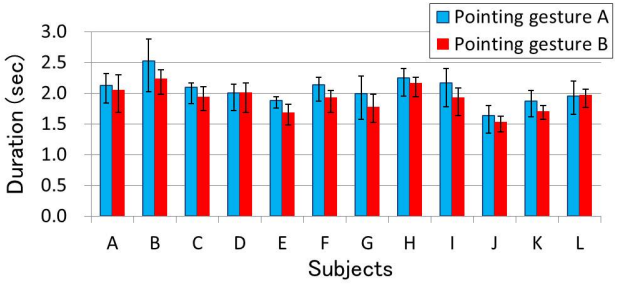

Figure 2: Duration of pointing gestures.

Akl, et al.[2] proposed gesture recognition method using a compressed sensing (CS) technique. CS is utilized to consider the effects due to differences in the data length. They focused on hand gestures that appear to be sparse since the hand follows a smooth trajectory during performing a gesture. Figure 3 is an example of a waveform for inertial sensor data of pointing gesture A, and Figure 4 is a result of Discrete Cosine Transform (DCT) for the data of Figure 3. These figures show that gesture traces can be represented using fewer samples as per the theory of compressive sensing[5].

For CS method, all the traces are projected into the same lower dimensional subspace and solve the problem of different durations. Simultaneously, the computational cost is reduced. Finally, the effect on the difference of duration is eliminated. As a result, Akl, et al. achieved user-independent recognition with high accuracy even when it is learned from training data of others[2].

It is not necessary for a new user to acquire training data if the user-independent recognition is realized. Thus, the burden on the user is reduced. However, the effectiveness of compressed sensing technique for the user dependency to pointing gesture recognition is unknown. Akl, et al. operated experiment with drawing the simple shape[2]. Figure 5 shows the acceleration data waveform during drawing the circle, and Figure 6 shows the data of Pointing gesture A. These figures show that these gestures are quite different. The pointing gesture is required to enforce by great force which is depending on the strength of each person. The changes of forces affect to the amplitude of the waveform which was not considered in existing work[2]. Therefore, in this paper, we tested the recognition accuracy for the pointing gesture using compressed sensing, and examined the training data reduction by user-independent recognition.

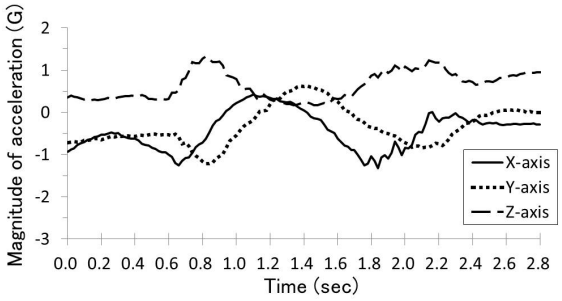

Figure 5: Sensor data for drawing a circle.

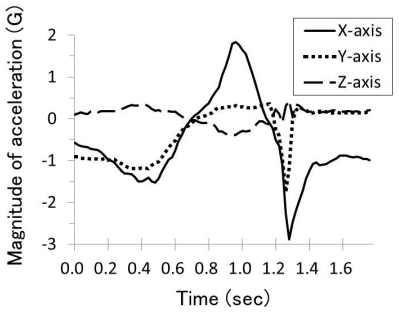

Figure 6: Sensor data for Pointing gesture A.

Pointing Gesture Recognition using Compressed Sensing

The implementation for gesture recognition system using CS followed the technique proposed by Akl, et al.[2]. The details of the implementation are explained in the next section.

Recognition Method

We assume that the data for recognition are obtained from inertial sensors. The sensor is attached to the wrist which is including 3-axis accelerometer and 3-axis gyroscope. Then, compressed template data matrix R of Figure 7 is generated using the acquired data. One column represents one training data. Also, compressed input data matrix Y is generated from input data. After the matrices are prepared, the system recognizes the gesture by solving the L1 norm optimization problem shown in equation (1). The lp_solve[13], a free library of Java, is used to solve the linear optimization problem.

$$\arg\min |\theta|_1 \text{ subject to } \mathcal{Y} = \mathcal{R}\theta \tag{1}$$

Each time series data for an axis of angular velocity and acceleration is used in equation (1). As a result of the equation (1), a θ is given. After θ for all axes are obtained, the absolute value of ith row of each θ are summed. Then, the sum is used to recognize the input data. The ith row of θ corresponds to the ith column of R. Moreover, the sum of absolute value increases when the similarity is high. Therefore, recognition of input data is realized by finding highest ith row of θ. Then, the system recognizes the input data as the same gesture of ith column of R.

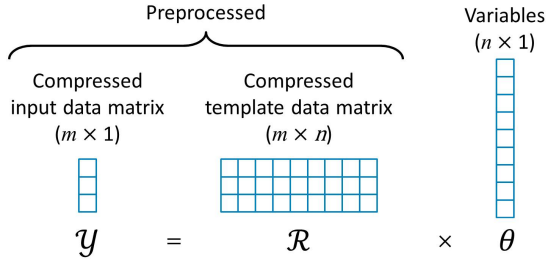

Figure 7: Equation of linear optimization.

Preprocessing

Before the recognition process described in the previous section, input data and template data matrix are prepared. More specifically, the impact on different data length must be considered as follows. First, the maximum data length l of the acquired input data and the training data are found. Input data are spotted to determine the data length for real use. Gesture data are spotted by such techniques like Stiefmeier, et al.[15] proposed. Then, zeros are added to the end of the data that has less length than l. After that, the data are stored in the template data matrix that has size of $\{$maximum length $l \times$ number of training data $n\}$. In addition, each column of the template data matrix has one repetition data of a gesture.

Then time sequence data are transformed into the sparse representation, where most elements are 0. We used discrete cosine transform. After the data are transformed into sparse data, input data matrix of $l \times 1$ and template data matrix of $l \times n$ are compressed from l dimensions to m dimensions using random matrix of $m \times l$. This process is called a random projection[5][4]. The random projection process reduces the influence of noise and the computational complexity. The m and l are related as $\{m < l\}$, and m is determined experimentally.

There are various proposed random matrix. However, we used simple solution that many researches successfully saved computational cost. The computational cost is saved since computations only using integer arithmetic[1]. The matrix is filled with rules of (2).

$$c_{ij} = \sqrt{3} \begin{cases} +1 \; with \; probability \; \frac{1}{6} \\ 0 \; with \; probability \; \frac{2}{3} \\ -1 \; with \; probability \; \frac{1}{6} \end{cases} . \quad (2)$$

Experiments

We conducted experiments to clarify the feasibility of recognition for the pointing gesture using CS. For actual use of the gesture recognition system, the other activities may be misrecognized as a pointing gesture. Thus, the experiments are required to investigate about misrecognition. Therefore, "pushing a switch", "handshake", and "walking" are included to recognition targets in addition to Pointing gesture A and B. During these activities, an arm starts from beside the body, then moves to the front of the body that is similar to the pointing gesture. Accordingly, in this paper, we recognize the following five activities.

(a) Pointing gesture A.
(b) Pointing gesture B.
(c) Pushing a switch.
(d) Handshake.
(e) Walking.

The data were acquired from 12 students (including 3 women). Each subject repeated 20 times for five gestures that are (a), (b), (c), (d), and (e). The inertial sensor (ATR-Promotions: WAA-010[3]) including 3-axis accelerometer and 3-axis gyroscope is mounted on the wrist and used to acquire the data. In addition, the acquisition frequency of inertial sensor was 50Hz. For noise reduction in high frequency domain, a sliding window is applied to act as a moving average filter with a width of 10 data samples.

To compare the recognition accuracy, we tested with DTW by same conditions. Also, the recognition accuracy is observed with two conditions which are user dependent (UD) and user independent (UI). For the UD condition, the template data contain training data acquired from an original subject. On the other hand, for the UI condition, the template data contain training data acquired from the other 11 subjects. Comparing the methods and conditions, precision and recall are computed. Our goal is to recognize the gesture (a) and (b). Therefore, the classification targets are 3 classes such that (a), (b), and the others, i.e. (c), (d), and (e). Also, F-measure represented by equation (3) is introduced to analyze the results.

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

The compression dimension is experimentally determined. We analyzed the relation between compression dimension and F-measure. The F-measure increased as compression dimension increased until F-measure gets to the peak. After the peak, the F-measure decreased while compression dimension is increased. Therefore, the compression dimension m is determined by finding a point where the F-measure is highest. As a result, m was 8 for UD condition which used training data of one subject. On the other hand, m was 21 for UI condition, which used training data of 11 subjects.

Results

Confusion matrices of the experimental results are shown in Tables 1 to 4. Also, Figure 8 shows average, maximum and minimum values of the average F-measure for classification of (a) and (b) in all subjects. For the user-dependent condition, UD-DTW averaged 0.80 and UD-CS averaged 0.95. Also, for the user-independent condition, the averages were 0.62 by UI-DTW and 0.79 by UI-CS.

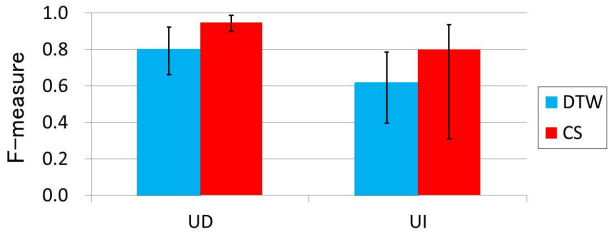

Figure 8: F-measure at each condition.

Discussion

User dependency

In Figure 8, the average of UD-DTW is no more than 0.8. While DTW is a strong gesture recognition method, this result suggests that the five gestures were similar. Also, there were differences between duration while repetition as we mentioned (see Figure 2). However, UD-CS is more than 0.9 even at the minimum value. Therefore, CS is considered as a robust method to the changes in a user.

The UI-CS is 0.18 higher than UI-DTW, even though we used the same data that varied in duration among users as shown in Figure 2. In addition, the difference between condition UD and UI for CS, which is the influence of the user dependency, is 0.02 less than DTW. These results suggest that CS is a better method compared to DTW for variances among users.

Figure 9 shows the F-measures of each subject at UD-CS and UI-CS. The reduction amounts from UD-CS to UI-CS are different by each subject. For some subjects, the reduction amounts are small. If the recognition accuracy for gestures of a person is high enough by using the training data of the others, training data is not necessary to acquire from his/her. Thus, the burden on a user is reduced. On the other hand, the reduction amounts are large for some subjects. Therefore, if the recognition for gestures of a person is not high enough by using the training data of the others, training data are necessary to acquire from his/her.

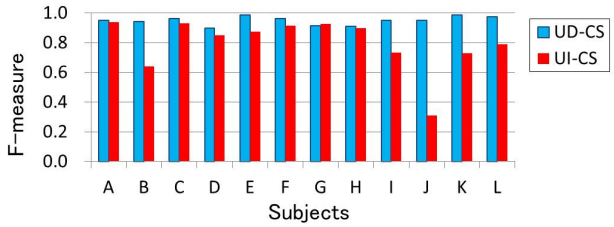

Figure 9: F-measure of user.

Acquisition of the training data

As mentioned above, user-dependency varies for each subject. Using this property, we tested the reduction in the number of users who require the acquisition of training data. The procedure for the test is shown in Figure 10. We investigate the recognition accuracy using training data of only specific users. The $order$ in Figure 10 is the order of priority for use of training data. These are determined in ascending order of the F-measure at UI-CS. For example, when u in Figure 10 is 2, test simulates a condition that only two subjects are required to acquire the training data. Therefore, for subject J and B, the training data include their own data in addition to the data of others. By contrast, for the rest of the subjects, the training data do not include data obtained from them.

```
Initialize order[12] by accuracy of UI-CS
 order[1] = Subject J
 order[2] = Subject B
 order[3] = Subject K
 …

Loop1: to all 12 conditions
 u is the number of users in template data
 u starts from 1, end after 12

 Loop2: to all acquired data
  Create testing data
   Pick one acquired data t for testing data

  Create training data
   Loop3: to all acquired data except t
    If acquired data x are
    acquired by order[0] , or ···, or order[u]
      Adapt data x to training data

  Recognize testing data using training data
```

Figure 10: Experimental procedure.

Figure 11 represents F-measures of the pointing gesture when performing classification under the conditions described above. Here, right next to the data is an identifier of the subject. These identifiers correspond to Figure 9. The test resulted in the same accuracy of UI-CS by using the training data of 4 subjects compared to 11 subjects at the UI-CS condition. In brief, 7 subjects are freed from data acquisition. Accordingly, the burden on the 7 subjects are released. Also, utilizing the compressed sensing method was found to be effective to use data of other subjects.

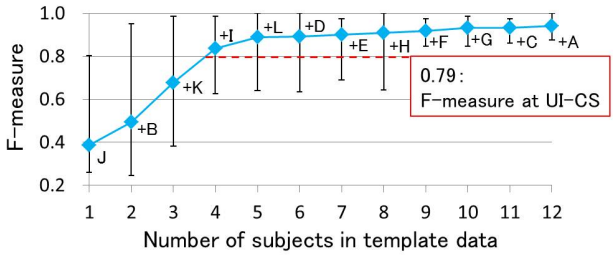

Figure 11: Number of training data by subjects.

Future work

Gesture recognition generally requires training data from original user[12][8]. In other words, the accuracy decreased by using the training data of others. In addition, several repetitions are necessary for each gesture[8]. Even though compressed sensing is utilized, the training data were required for specific users.

However, our results suggested that utilizing the compressed sensing is effective to reduce the number of subjects who are required for training data acquisition. Also, identifying the users who need original training data was effective. Therefore, the next challenge is the determination of the specific user who is high in the user-dependency. An idea is that a recognition accuracy against an existing recognition model is evaluated to see if it is high enough to be considered as "standard user", i.e. user-independent. Here, we assume that a small amount of data are collected from all users prior to the evaluation.

Conclusion

In this paper, we investigated about the recognition of the pointing gesture during pointing and calling using the compressed sensing. Also, we discussed training data reduction through the use of data obtained from others. As a result of the experiment, compared with the conventional method using DTW, F-measure was improved 0.18 even the data obtained from others are used. From these results, we found that the recognition using compressed sensing is considered as a robust method to the changes in a user. Furthermore, gesture recognition using compressed sensing is found to be robust to changes among users.

Furthermore, we found that the user-dependency varies for each subject. Thus, we tested to recognize the pointing gesture of all subjects by using the training data of only specific users. The specific users are determined in the order of user-dependency for each subject. The test showed that the recognition model with training data from 4 specific subjects provided the same level of accuracy as the one from 11 subjects. This suggested the feasibility of reduction for subjects who need to acquire the training data. For future work, we will investigate a method to determine the user dependency for each user by using a small number of input data.

Acknowledgment

This work is partially supported by MEXT Grants-in-Aid for Scientific Research (A) No. 23240014 and (C) No. 24500142.

Table 1: UD-DTW, F-measure:0.77

		Output					
		a	b	c	d	e	Recall
Input	a	**137**	0	35	34	34	57.1%
	b	0	**239**	1	0	0	99.6%
	c	26	1	**187**	16	10	77.9%
	d	30	0	25	**164**	21	68.3%
	e	17	0	1	5	**97**	82.1%
Precision		65.2%	99.6%	72.2%	71.6%	75.2%	77.0%

Table 2: UD-CS, F-measure:0.94

		Output					
		a	b	c	d	e	Recall
Input	a	**225**	7	2	4	2	93.8%
	b	6	**227**	2	3	2	94.6%
	c	0	0	**235**	4	1	97.9%
	d	4	0	25	**205**	6	85.4%
	e	0	3	1	1	**235**	97.9%
Precision		95.7%	95.8%	88.7%	94.5%	95.5%	93.9%

Table 3: UI-DTW, F-measure:0.58

		Output					
		a	b	c	d	e	Recall
Input	a	**57**	0	55	73	55	23.8%
	b	0	**237**	3	0	0	98.8%
	c	41	1	**151**	35	12	62.9%
	d	58	16	38	**97**	31	40.4%
	e	22	1	27	25	**165**	68.8%
Precision		32.0%	92.9%	55.1%	42.2%	62.7%	58.9%

Table 4: UI-CS, F-measure:0.82

		Output					
		a	b	c	d	e	Recall
Input	a	**191**	34	5	3	7	79.6%
	b	38	**191**	1	7	3	79.6%
	c	1	0	**204**	34	1	85.0%
	d	12	2	57	**168**	1	70.0%
	e	3	1	7	2	**227**	94.6%
Precision		78.0%	83.8%	74.5%	78.5%	95.0%	81.8%

References

[1] Achlioptas, D. Database-friendly random projections. In *Proc.PODS*. (2001), 274-281.

[2] Akl, A., Feng, C., and Valaee, S. A Novel Accelerometer-Based Gesture Recognition System. *IEEE Transactions on Signal Processing*, (2011), 6197-6205.

[3] ATR-Promotions: WAA-010. http://www.atr-p.com/sensor10.html.

[4] Bingham, E., Mannila, H. Random projection in dimensionality reduction: Applications to image and text data. In *Proc. KDD*, (2001), 245-250.

[5] Candes, E., Wakin, M. An introduction to compressive sampling. *IEEE Signal Process. Mag.*, (2008), 21-30.

[6] Concept of Zero-accident Total Participation Campaign. http://www.jniosh.go.jp/icpro/jicosh-old/english/zero-sai/eng/index.html.

[7] Eickeler, S., Kosmala, A., and Rigoll, G. Hidden Markov Model Basaed Continuous Online Gesture Recognition. In *Proc. ICPR*, (1998), 1206-1208.

[8] Gillian, N., Knapp, R.B., and O'Modhrain, S. Recognition of multivariate temporal musical gestures using n-dimensional dynamic time warping. In *Proc. NIME*, (2011), 343-348.

[9] Iwasaki, M., Fujinami, K. Recognition of Pointing and Calling for Industrial Safety Management. In *Proc. ICT-ISPC*, (2012), 50-53.

[10] Jelinek, F. Continuous speech recognition by statistical methods, In *Proc. the IEEE*, (1976) 532-556.

[11] Lee, H., Kim, J.H. An HMM-based threshold model approach for gesture recognition. In *TPAMI*, (1999), 961-973.

[12] Liu, J., Zhong, L., Wickramasuriya, J., et al. UWave: Accelerometer-based personalized gesture recognition and its applications. In *Proc. PMC*, (2009), 657-675.

[13] lp_solve. http://lpsolve.sourceforge.net/5.0/.

[14] Mohamed, A., Ramachandran, K.N.N. Continuous Malayalam speech recognition using Hidden Markov Models. In *Proc. A2CWiC*, (2010) 532-556.

[15] Stiefmeier, T., Roggen, D., and Troster, G. Gestures are strings: efficient online gesture spotting and classification using string matching. In *Proc. BodyNets*, (2007), 1-8.