Gesture recognition based on 3D accelerometer for cell phones interaction

Zhenyu He, Lianwen Jin*
School of Electronic and Information Engineering,
South China University of Technology,
Guangzhou, China.510640
*E-mail: lianwen.jin@gmail.com

Lixin Zhen, Jiancheng Huang
Motorola China Research Center,
Shanghai, China, 210002

Abstract—**This paper proposes a gesture recognition system based on single tri-axis accelerometer mounted on a cell phone for human computer interaction (HCI). Three feature extraction methods, namely discrete cosine transform (DCT), Fast Fourier transform (FFT) and a hybrid approach which combine wavelet packet decomposition (WPD) with Fast Fourier transform are proposed. Recognition of the gestures is performed with Support Vector Machine (SVM). Recognition results are based on acceleration data collect from 67 subjects. The best average recognition result (87.36%) for 17 complex gestures is achieved with wavelet-based method, while DCT and FFT produce accuracy of 85.16% and 86.92% respectively. The performance of experimental results shows that gesture-based interaction can be used as a novel HCI for mobile applications, such as games control and music navigation.**

I. INTRODUCTION

Context awareness is an emerging application area with the aim of easing human computer interaction (HCI). In the case of a mobile device the HCI can be tedious given the physical size limitation both in the keyboard and screen [1]. If the mobile terminal can aware of the user's current context then it could react in some appropriate manner to suit the user without the need of user interaction.

Since gestures are commonly used in daily life, gesture-based interaction can be one of novel interaction ways that users want. To implement the gesture-based interaction, many different techniques, such as vision-based gesture interaction, touch-based gesture interaction have been utilized [2]. In recent years, a new kind of interaction technology that recognizes users' movement has emerged due to the rapid development of sensor technology. An accelerometer measures the amount of acceleration of a device in motion. Analysis of acceleration signals enables three kinds of gesture interaction methods: tilt detection, shake detection and gesture recognition [2-5].

Although in the literature there are already exist some approaches of using acceleration signals for gestures recognition, most work focuses on recognizing the simple gestures such as Arabic numerals [2-4], simple linear movements and direction [5]. In our work, we attempt to recognize 17 complex gestures from 67 volunteers' acceleration data.

As activity recognition can be formulated as a typical classification problem and just like many pattern recognition problem, features extraction plays a crucial role during the recognition process. However, few works that extract effective features and make quantitative comparison of their quality are reported. To extract feature from the acceleration data, they convert three dimensional data into one dimensional vector using vector quantization [5]. Some work use acceleration, velocity, position and combination of acceleration with velocity respectively to recognize 10 Arabic numerals [4]. Others work extracts the statistics of acceleration data such as local maximal or minimal point as feature [3]. Besides, acceleration signals are sampled in equal-time interval, the length of data is variable according to the different gesture and different subject's input speed. Therefore, some present works adopt Dynamic Time Warping or Hidden Markov Model algorithm for gesture recognition [2, 4, 5]. However, some other better recognize algorithms such as Support Vector Machine (SVM) [6] cannot be used unless we firstly resample the acceleration data into same number of point [3]. Instead of resampling the data, we adapt discrete cosine transform (DCT) and Fast Fourier transform (FFT) to extract the primary information of data and reduce the dimensions of data.

In this paper, three feature extraction methods, namely discrete cosine transform, Fast Fourier transform and a hybrid approach which combine wavelet packet decomposition (WPD) with FFT are proposed. Experimental results show that all three proposed feature can recognize the 17 complex gestures base on single tri-accelerometer. The average recognition results for DCT, FFT and wavelet-based are 85.16%, 86.92% and 87.36% respectively.

The remainder of this paper is organized as follows. In section Ⅱ, we introduce the data collection. Section Ⅲ presents the detailed information about the feature extraction, including DCT, FFT and WPD. Section Ⅳ introduce classification method and experiment results is given in section Ⅴ. Finally, conclusions are given in section Ⅵ

This paper is supported by Motorola Research Found (B06-D8070740) and partially supported NSFC (No. U0735004, 60772216).

978-1-4244-2342-2/08/$25.00 ©2008 IEEE.

II. DATA COLLECTION

As shown in Figure 1(a), a single tri-axis accelerometer is mounted on a cell phone to collect different gestures data. Sixty-seven subjects held the cell phone in hand and performed 17 different gestures in different days. The exact sequence of gestures is listed in Table 1. Each subject performed every gesture once. The output signal of the accelerometer is sampled at 300 Hz. Since acceleration signals are sampled in equal-time interval, the length of raw data is variable according to different gesture and different input speed. The longest length of collected data is 12810 sampled points while the shortest length of data is 1270 sampled points. Data from the accelerometer has the following attributes: time, acceleration along X-axis, Y-axis and Z-axis. Figure 1 (b) shows the example of raw data.

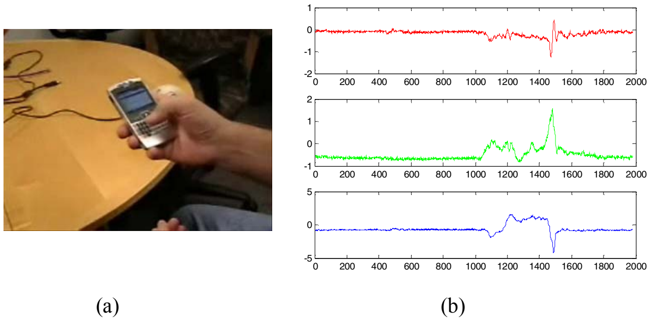


Figure 1. (a) Setup of data collection and (b) example of raw data

TABLE I. GESTURES LABELS

Class	Gestures
1	Tilt phone to left & back, then to right & back (>15 deg)
2	Tilt phone towards & then away from you (>15 deg)
3	Slowly tilt phone 90 deg to the left & back, then to right & back
4	Slowly tilt phone 90 deg towards & then away from you
5	Shake phone with no specific direction once
6	Shake phone to the left & back, then to right & back
7	Shake phone towards you & back, then away from you & back
8	Pan phone upward & downward & right & left
9	Tap phone on top left & right, then bottom left & right corner
10	Pick up phone from table, hold to view, & back to table
11	Pick up phone from table & bring it to ear & back to table
12	Bring phone from holding for viewing to ear & back to viewing
13	Take phone off belt clip & hold & put it back
14	Phone in the pocket (no intentional motion)
15	Rotate phone from portrait to landscape & back to portrait
16	Roll phone to left & back, then to right & back
17	Move phone towards, then away from your face

III. FEATURE EXTRACTION

Figure 2 shows an overview of the proposed framework for gesture interactive. When a user performs gestures on 3D space using the mobile phone, the movement is sensed by an accelerometer. Then the acquired data is processed and classified into a gesture through the gesture recognition algorithm. Finally, the corresponding function is executed and feedback to the users.

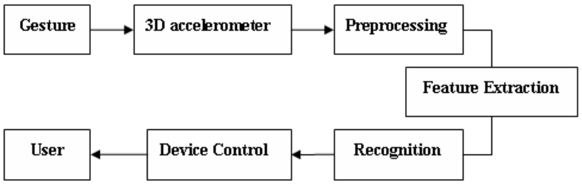


Figure 2. Framework of Gesture Interactive for Cell Phone

Feature extraction is the elementary problem in the area of pattern recognition. For gesture recognition task, extraction of effective gesture features is a very important step which will greatly improve the performance of the gesture recognition system. Therefore, we proposed three effective features extraction methods from acceleration data in this paper and the details of these methods are presented as follows.

A. DISCRETE COSINE TRANSFORM

Discrete cosine transform (DCT) [7] is an important orthogonal transform and its performance has been found to be asymptotically equivalent to the optimal Karhunent-Loeve transform for signal decorrelation. The DCT of a data sequence $x(n), n = 0,1,\cdots,(N-1)$ is defined as:

$$X_c(0) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) \qquad (1)$$

$$X_c(k) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x(n) \cos\frac{(2n+1)k\pi}{2N}, k = 1,2,\cdots,(M-1) \qquad (2)$$

Where $X_c(k)$ is the kth DCT coefficient. All N DCT coefficients can be computed using a 2N-point fast Fourier transform. It can be shown that $X_c(k)$ is a bandpass filter with a center frequency at $(2k+1)/\ 2N$ when the sampling frequency is normalized to 1. Hence, the magnitude of the output of $X_c(k)$ for small k is generally larger. In other words, the DCT can be concentrated in the low indices of the DCT if the remaining DCT coefficients can be set to zero without a significant impact on the energy of the signal. DCT is widely used in image compression also because of its excellent energy compaction property. As shown in Fig. 3, lots of frequency component of our gesture acceleration data are centralized at the low-frequency. Most of the visually significant information is concentrated in just a few DCT coefficients too. Therefore, we discard the high-frequency DCT coefficients, and select the low-frequency DCT coefficients as gestures features. In this paper, we extract the first 128 magnitude of DCT coefficients from each axis acceleration data for features.

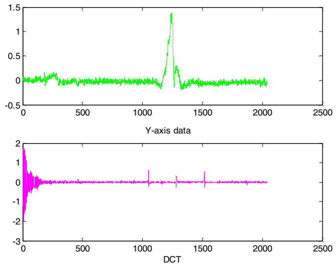


Figure 3. Y-axis data and its DCT

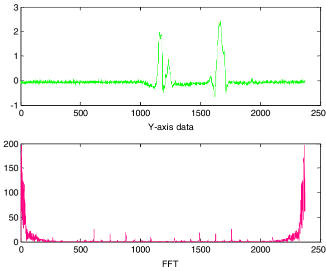


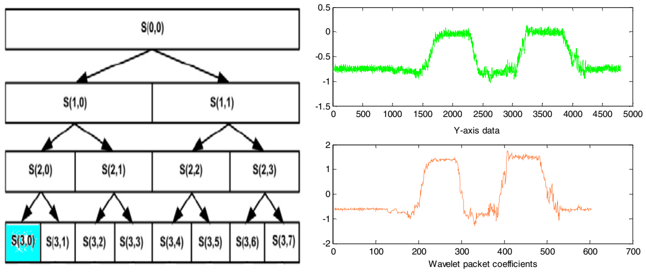
Figure 4. Y-axis data and its FFT

B. *Fast Fourier transform*

Fast Fourier transform (FFT), is a typical signal processing approach which can be used to transform the signal from spatial domain to frequency domain. Figure 4 is an example of Y-axis accelerometer data and its FFT coefficients. Similar to DCT, lots of frequency component are centralized at the low-frequency too. Therefore, we also extract the first 128 magnitude of FFT coefficients from each axis acceleration data as features. Our experiment has shown that using these low-frequency features not only hold the primary information, but also reduce the dimensions of data.

C. *Wavelet packet decomposition*

In order to analyze acceleration data more accurately, we employ wavelet packet decomposition (WPD) [8]. It works by generalizing the link between multiresolution approximation and wavelet bases. Compared to wavelet transform, WPD not only decomposes the approximation coefficients, but also the detail coefficients [9]. A WPD is shown in Fig. 5 (a), where $s(0,0)$ denotes the original signal space, $s(j,k)$ denotes the decomposed subspace, *j* is the decomposition level, and *k* is the index of the subspace occurring at the *j*th level. Therefore, wavelet packet decomposition can decompose the signals to different frequency range ideally. By using WPD, we can obtain decomposed signals that can efficiently represent the features of signal patterns.



(a) (b)

Figure 5. (a) Structure of WPD and (b) WP coefficients of node (3,0)

As discuss above, changes in gestures are characterized mainly by low-frequency of signal. Thus, the low-frequency component, which includes gesture information, can discriminate the different gestures efficiently. Therefore, we decompose the original signals three level using Daubechies wavelet of order 3 and then we obtain wavelet packet coefficients of node s(3,0) which represents the low-frequency of signal. Figure 5 (b) shows an example of Y-axis accelerometer data and its wavelet packet coefficients of node s(3,0). It can be seen that wavelet packet decomposition not only extracts the primary information effectively, but also remove the high-frequency noise and random dithering of signal. After that, we transform wavelet packet coefficients using FFT and extract the firs 128 FFT magnitude of coefficient as features.

IV. Classification Method

The classification algorithm we used is Support Vector Machine (SVM) [6]. We used One-versus-One Strategy (OVO), where a set of binary classifiers are constructed using corresponding data from two classes. While testing, we used the voting strategy of "Max-Wins" to produce the output.

Five-fold cross-validation was used for classifier assessment. The data was randomly divided into five groups with the same number of samples for different classes. The classifier was built five times. Each time one group in turn was excluded from the training and used solely as a test set. The cross-validated classification result is the average of the five testing results.

V. Experimental Results

In the experiments, we carried out five-cross-validation procedure to validate the effectiveness of the proposed features. The recognition results of DCT coefficients，FFT coefficients and combine WPD with FFT are given in Table II.

TABLE II. Recognition Accuraacy of Three Proposed Features

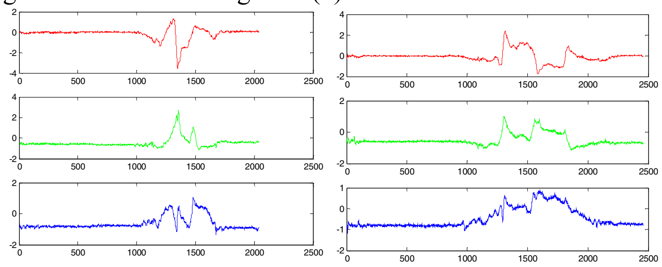
Class	**Accuracy**		
	DCT	***FFT***	***WPD+FFT***
1	89.56	86.59	89.56
2	92.53	92.53	92.53
3	79.34	82.20	82.20
4	83.41	87.91	87.91
5	64.51	70.44	71.98
6	67.47	80.55	82.09
7	86.59	88.24	88.24
8	85.27	85.38	85.38
9	94.18	94.29	94.29
10	92.64	92.64	92.64
11	80.77	85.16	85.16
12	92.53	89.67	89.67
13	82.42	85.05	85.05
14	83.63	81.98	81.98
15	91.21	94.06	94.06
16	95.49	94.06	95.49
17	88.46	88.35	88.35
Average	**85.16**	**86.92**	**87.36**

TABLE III THE CONFUSION MATRIX FOR WAVELET-BASE FEATURES

class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	60		4					1	1		1						
2		62		2						1		1					1
3	7		55	1							1				3		
4	1	7		59													
5		1			48	**7**	**8**		1	1		1					
6	1		1		**9**	55					1						
7		1		1	2	1	59	1	1				1				
8		1			2	1		57	1		1	1					3
9						1			63						1		2
10							1		2	62		1		1			
11								2			57	4	1	3			
12								3	1		2	60		1			
13								1		1	3		57	4		1	
14										2	3	5	1	55		1	
15			1						2	1					63		
16								1					2			64	
17		2						1	4			1					59

Table II show that all three proposed features can recognize the 17 complex gestures based on single tri-accelerometer. Particularly, the wavelet-based method outperforms the others while the performance of using DCT coefficients is only slightly lower. The average recognition results for DCT, FFT and wavelet-based are 85.16%, 86.92% and 87.36% respectively. Experimental results show that using DCT and FFT not only hold the primary information, but also reduce the dimensions of data. By using WPD, we can obtain decomposed signals that can efficiently represent the features of signal patterns.

In order to find out which gestures are relatively harder to be recognized, we analyzed the confusion matrices. Table III shows the aggregate confusion matrix for our wavelet-based features. It can be seen that the fifth gesture (Shake phone with no specific direction once) is hard to recognize. Because the fifth gesture often confuse with the sixth gesture (Shake phone to the left & back, then to right & back) and the seventh gesture (Shake phone towards you & back, then away from you & back). This result is reasonable, because the raw signals of the fifth gesture are similar to the sixth gesture and the seventh gesture. An example of the fifth and sixth gesture is shown in Figure 6 while the raw signal of the seventh gesture is shown in Figure 1 (b).



(a) (b)

Figure 6. Raw data of the fifth gesture (a) and the sixth gesture (b)

VI. CONCLUSION

In this paper, we propose a gesture recognition system based on a single tri-axis accelerometer mounted on a cell phone for human computer interaction. Three feature extract method, namely discrete cosine transform, Fast Fourier transform and combine wavelet packet decomposition with Fast Fourier transform are proposed. Classification of the gestures is performed with Support Vector Machine. Gesture recognition results are based on acceleration data collect from 67 subjects. The best recognize rate (87.36%) is achieved with wavelet-based method, while DCT coefficients and FFT coefficients produced accuracy of 85.16% and 86.92% respectively. Experimental results show that using DCT and FFT not only hold the primary information, but also reduce the dimensions of data. By using WPD, we can obtain decomposed signals that can efficiently represent the features of signal patterns. Gesture recognition based on single tri-axis accelerometer mounted on a cell phone provides a novel human computer interaction.

REFERENCES

[1] J.A.Flanagan and J.Mantyjarvi, “Unsuperised clustering of symblo strings and context recogniton”.ICDM, Maebashi,Janpan. pp.171-178 .

[2] Eun-Seok Choi, Won-Chul Bang, et. al, “Beatbox Music Phone: Gesture Interactive Cell phone using Tri-axis Accelerometer,” ICIT IEEE Int. Conference on Industrial Technology, 2005.

[3] Sung-Jung Cho, Eunseok Choi, et. al., “Two-stage Recognition of Raw Acceleration Signals for 3-D Gesture- Understanding Cell Phones”, 10th IWFHR, La Baule, France, Oct. 2006.

[4] Sung-Do Choi, A.S. Lee, Soo-Young Lee, “On-Line Handwritten Character Recognition with 3D Accelerometer”, 2006 IEEE International Conference on Information Acquisition, pp.845-850,2006.

[5] S. Kallio, J. Kela and J.Mantyjarvi, “Online gesture recognition system for mobile interaction”, 2003 IEEE International Conference on Systems, Man and Cybernetics, vol 3, pp.2070-2076.

[6] V. Vapnik. The nature of statistical learning theory, Springer Press, New York, 1999.

[7] N. Ahmed, T. Natarajan, and K. R. Rao. “Discrete Cosine Transform”, IEEE Trans. on Computers, vol. 23, pp.90-93,Jan 1974.

[8] R. R. Coifman, Y. Meyer, and M. V. Wickerhauser, "Wavelet analysis and signal processing," in Wavelets and Their Applications, M. B. Ruskai, Ed. Boston: Jones and Bartlett, 1992.

[9] L. Deqiang, W. Pedrycz, and N. J. Pizzi, “Fuzzy wavelet packet based feature extraction method and its application to biomedical signal classification”, IEEE Transactions on Biomedical Engineering, vol. 52, pp.1132-1139,2005.