Hand gesture recognition of a mobile device user

Vesa-Matti Mäntylä[1], Jani Mäntyjärvi[2], Tapio Seppänen[2], Esa Tuulari[1]

[1]Technical Research Centre of Finland, Oulu, Finland
[2]Department of Electrical Engineering, University of Oulu, Oulu, Finland

ABSTRACT

Experiments with acceleration sensors is described for static and dynamic gesture recognition of a mobile device user. Static gestures are recognized with the self-organizing mapping scheme of Kohonen while a hidden Markov model is used for recognizing dynamic gestures. An experimental sensor box for the research of context-awareness is also explained. Experimental results show great promise of the chosen technologies for recognizing gestures of a user of a mobile device.

1. INTRODUCTION

The needs of developing systems for human/computer or human/robot interfaces has lead to a number of succesful attempts of using hand gesture recognition. The most conventional approaches to hand gesture recognition have employed cybergloves[8]. The computer vision community has also shown a lot of interest recently in the recognition of human actions and gestures [7], [11], [12], [13]. In the work of Yang et al. [21] a mouse was used as a two-dimensional gesture input device. The human-robot interface was developed for robot teleoperation and programming by Lee et al. [8]. This and a human-computer interface have also been proposed as an application for the system of Yang et al. [21].

The accelerometer-based arm gesture recognition system of Harrington et al. [3] was studied as an alternative method of computer input for people with severe speech and motor impairment. Musical performance control and conducting recognition systems have been created by Sawada et al. [18] and also Usa et al. [20] Accelerometer-based measurement of gestures has been carried out succesfully in the works of Harrington et al. [3], Sawada and Hashimoto[18] and Usa et al. [19].

Regarding gestures as temporal feature trajectories with temporal and spatial variations has lead to the use of hidden Markov models (HMMs), succesfully applied in speech recognition, for example by Levinson et al.[9]. HMM has been used e.g. in the works of Lee et al. [8], Min et al. [12], Yang et al. [21] and also Usa et al. [20]. The most commonly used topology of HMMs has been left-right topology [9], which is exclusively used in the previously mentioned systems involving hand gesture recognition.

We have built a general purpose sensor-box for studying context-awareness [19]. The box includes several types of sensors of which only acceleration sensors were used in this study. The accelerometers measure both dynamic accelerations like vibrations and static accelerations like gravity. The box was attached to a mobile phone in order to deduce automatically in which state the user has the phone. A set of commonly occuring hand gestures of a mobile device user was selected and further divided into two categories. Static gestures include still poses of the sensor box, which occur, e.g., when keeping the phone on ones ear while talking to it. A dynamic gesture is related to a movement of the sensor box which occurs, e.g., when moving the phone to ones ear in order to answer a call[1].

In this paper we present the results that we have obtained by using HMM and Self Organizing Map of Kohonen (SOM) in phone-gesture recognition. Both static poses and dynamic movements can be recognized. Chapter 2 present methods that we have used. Experiments and results are presented in Chapter 3, and Chapter 4 summarizes the work.

2. PHONE-GESTURE RECOGNITION METHODS

2.1 Static phone-gesture recognition with SOM

The self-organizing map of Kohonen [6] has been widely used in pattern recognition and feature extraction [2],[4],[5],[17]. In this work SOM was used to extract certain positions, i.e. static gestures, of the mobile phone. The accelerometers employed are capable of measuring both dynamic and static acceleration. The sensor box implemented into the mobile phone detects the 3D-acceleration of the mobile phone. By processing normal usage acceleration signals it is possible to define and extract certain position chains representing how the mobile phone is used. The positions of the mobile phone used in this study are 'display up', 'display down', 'in the pocket', 'on the left ear' and 'on the right ear'.

The SOM is a neural network, which forms spatially organized feature maps from n-dimensional input signal in an unsupervised manner. The method is similar to human sensory input mapping in the brain, which is then organized topographically. During training the weight vectors $\mathbf{w}_{ji}(n)$ of the network are shifted closer to the input vectors $\mathbf{x}$ by

$$\mathbf{w}_{ji}(n+1) = \mathbf{w}_{ji}(n) + \eta(n)n_{j,i(\mathbf{x})}(n)(\mathbf{x} - \mathbf{w}_{ji}(n)), \quad (1)$$

where $\mathbf{w}_{ji}(n+1)$ is the updated weight vector, $\eta(n)$ is the learning rate and $n_{j,i(\mathbf{x})}(n)$ is the neighborhood function. The learning rate and the neighborhood function are changed as the training proceeds. After self-organization the hand-labeled training data are presented again to the SOM and labels are suggested to the winning neurons with their immediate neighbors. All suggestions are collected for each neuron after which majority voting is performed for final labeling.

[1] Patent Pending

0-7803-6536-4/00/$10.00 (c) 2000 IEEE

The static phone-gesture recognition system is divided into pre- and postprocessing modules. The preprocessing module includes anti-aliasing filtering and normalization, while the postprocessing module consists of feature extraction and classification. A block diagram of the static gesture recognition system of the mobile phone is presented in Figure 1.

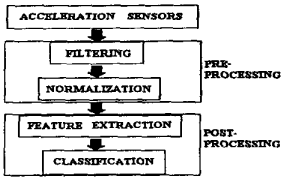


Figure 1. Diagram of the static gesture recognition system.

The SOM used here was a 2-dimensional hexagonal grid of size 7×7. Parameters of the SOM are presented in Table 1.

Table 1. Parameters of the SOM.

Parameter	Value/Argument
Map lattice	Hexagonal
Initialization type	Linear
Size of the map	7x7
Initial and final learning radiuses	[3 1]
Initial learning rate	0.03
Learning rate	Linear
Neighbourhood function	Epanechnikov

Acceleration signals are sampled with 90 Hz sampling frequency. The input to the preprocessing system is a sequence of vectors

$$\mathbf{P}_n = [x_n\ y_n\ z_n]^T, \qquad (2)$$

where x_n ,y_n ,z_n are acceleration signals in x, y, and z directions at discrete time n. In the preprocessing block each vector component is filtered and normalized separately. A low-pass filtering is carried out with a fourth-order Butterworth filter of type IIR. A 3-dB cut-off frequency of 2.5 Hz was considered suitable for static gestures. The variance of each component is normalized to 1, and the mean to 0. For the postprocessing block the components are finally recombined to a sequence of 3-dimensional feature vectors $\mathbf{P'}_n$. A feature vector $\mathbf{P'}_n$ at time n is input to a two-dimensional feature map (SOM codebook) which then produces an index I_n representing the training data cluster in question. The resulting sequence of SOM indices I_n is interpreted by the classifier, which assigns a label to each index according to the training data. Each label suggests a specific gesture. Finally, a majority voting is performed among the label sequence to recognize the gesture.

2.2 Dynamic phone-gesture recognition with HMM

The hidden Markov model (HMM) is a doubly stochastic process with an underlying process of transitions between hidden states of the system and a process of emitting observable outputs. When the outputs are discrete symbols, we talk about discrete HMM. The state transitions form a first order discrete Markov process with a transition probability distribution A and an initial state distribution π. The observable process of emitting symbols can be presented as an observation symbol distribution B. Thus each HMM can be presented as a triple, $\lambda = (A,B,\pi)$.

In the application of the HMM, we face the basic problems of training and testing the HMM models. In this work, we have used Baum-Welch [1] and Viterbi algorithms for the training and recognition tasks, see e.g. [16]. To be accurate, we used the log-Viterbi form of the Viterbi algorithm, for well-known computational savings [16]. Before training, initialization of the HMM parameters was done as follows:

- the initial state probability for the first state was set to 1,
- transition probability distribution for each state was set uniformly distributed, and
- any topologically allowable state transition from individual states, was given probabilities of form *1/(*amount of allowable state-transitions from states).

A HMM with a left-right topology is often used for modeling time-series whose properties change sequentially over time. Left-right HMMs with seven states were used also in our work. As a result of varying the amount of states, we found out that this was not such a critical factor in the recognition process. Choosing five states for each model did not have any remarkable effect on the recognition ability of the system.

The collection of the three dimensional acceleration data is performed with 100 Hz sampling rate. A similar preprocessing is required as in the SOM case above. The following steps are performed after sampling and digitation:

1. filtering of each acceleration component with fourth-order lowpass Butterworth filter with the 3 dB cut-off frequency of 4 Hz,
2. decimation of the filtered signal at 1/6 times the original sample rate,
3. detection of the useful signal,
4. for individual gesture, normalization of each component to zero-mean and unit-variance.

A software package for HMM computation was implemented in our earlier project on speech recognition [14],[15]. Discrete codebook indices correspond to the observable symbols and are input to the HMM both in training phase and test phase. The indeces are computed by vector quantization of 3-D input vectors of the acceleration signal.

The codebook was constructed by quantizing uniformly the 3-D feature space. Uniform quantization is sensible because the acceleration trajectory in the feature space can pass through any point with equal probability within the region defined by the application. Minimum and maximum value of each acceleration component was searched in the measured training data in order to define the parallelogram of activity in the feature space. The size of codebook is 512 3-D codewords. Size was found critical a fact in the classification capability of the recognition system. When we made the codebook radically smaller than 512 codevectors, for example 256 codevectors, we came up with distinctly poorer classification results. This must also partly be due to the fact that this codebook was not a result of any clustering process of some training data. Vector quantization of input vectors during recognition is performed in a conventional way by selecting the codebook entry containing the closest codeword to the input vector in the Euclidian sense.

3. EXPERIMENTS AND RESULTS

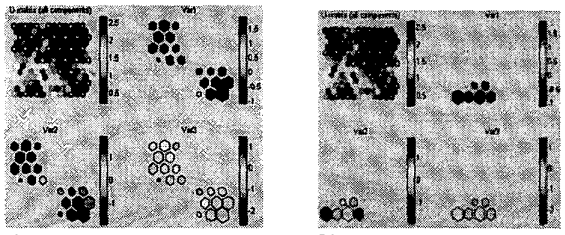
3.1 Experimental sensor box

In the sensor-box 3-dimensional acceleration information is achieved by combining the signals from one 2-D and one 1-D accelerometers. The accelerometers of type ADXL202 are manufactured by Analog Devices. The analog signal from the accelerometers is A/D-converted and sampled with National Instruments DaqCard 1200 measurement board which is connected to a PC-CARD slot in a laptop PC. The measurement program that stores the digitized acceleration signals on a disk is programmed in LabView which is a graphical programming environment from National Instruments. The quality of the signals can be checked visually with the measurement program before starting the recording. Acceleration signals are processed off-line in order to recognize static and dynamic gestures.

3.2 Static gestures

The training and test data were recorded in normal usage situations on six testees. The testees were asked to use the mobile phone as they normally did in the described usage situation and the acceleration data sequences were recorded. The static gestures were:
(SG1)'display up': Holding the phone in front of the chest and watching the display. Situation was recorded when testees were stationary and moving.
(SG2)'display down': The phone on the table display down.
(SG3)'on the left/right ear': The phone on the left /right ear. Testees were moving.
(SG4)'in the pocket': The phone in the left or right pocket. Testees were moving.

The amount of the position data from every situation was 4050 samples. The SOM feature map was trained with the test data of six testees which was decimated by the factor six. Figure 2 shows two examples of the result of self-organization. The sub-windows of var1, var2 and var3 display the three components of each code word. The size of each hexagon denotes the height of the histogram in each neuron and the color denotes the weight of the neuron. Only those BMU's are shown in which the perceptual values of hits of the labeling data were <0.01. The U matrix shows the cluster structure of the map. The deep colour areas in the U matrix compose 'valleys', which are clusters of different statics gestures.



a) b)

Figure 2. a) Clustering of the training data 'on the ear' in two different clusters 'on the right ear' and 'on the left ear'. b) Clustering of the training data 'in the pocket'.

This can be identified in Figure 2a) two clusters corresponding to the cases of 'on the left ear' and 'on the right ear'.The recognition accuracy in the 'display up' situation is notably lower than in other situations. This is due to the fact that the testees had a tendency of lifting the phone into upright position when watching the display, while the trainees kept the phone in a horizontal pose. A better performance is very likely to be gained by complement-ing the training material with samples covering these variations. The results of the recognition of static gestures are presented in Table 2.

Table 2. The results of static gesture recognition.

Test position	Recognition accuracy
Display up	89,4%
On the left ear	100%
On the right ear	96,6%
In the pocket	96,3%
Display down	100%

3.3 Dynamic gestures

In this work we have used six natural phone gestures related to the usage of a phone:
(DG1) An alarming phone, which is in right-hand side belt box with display turned towards the user, is taken into the right hand and the number of the caller is checked by watching. After this the phone is returned into the belt box in the same position.
(DG2) An alarming phone, which is in the right-hand side belt box with display turned towards the user, is taken into the right hand and the number of the caller is checked by watching. After this the user opens the line by pressing corresponding button on the console and brings the phone to the right ear.
(DG3) Inverse gesture of DG2.
(DG4) An alarming phone, which is on the desk with its display facing up, is taken into right hand and the number of the caller is checked by watching. After this the phone is put back on the desk in the same position.
(DG5) An alarming phone, which is on the desk with its display facing up is taken into right hand and the number of the caller is checked by watching. After this the user opens the line by pressing corresponding button on the console and moves the phone to the right ear.
(DG6) Inverse gesture of DG5.

The training data were collected from one person (P1) containing 100 repetitions of each gesture. The same person performed 50 repetions per gesture for testing. Another 50 repetitions, for testing, were collected from another person (P2). Experimental results are shown in Table 3.

It can be seen from Table 3 that the single-user case is handled with no errors, and the generalization of the recognizer to two users has been quite succesfull.

Table 3. The results of dynamic gesture recognition.

Test gesture	Recognition accuracy [%], (P1)	Recognition accuracy [%], (P2)
DG1	100	94
DG2	98	96
DG3	100	98
DG4	100	100
DG5	100	88
DG6	98	100

There are a number of factors causing errors in the two-user case:

1. dynamical differences (intensive <-->flegmatic),
2. temporal differences (slow <--> fast),
3. physical dimensions of the testee (length of the body and reach of the right hand),
4. standing pose of the testee,
5. initial, intermediate and final phone positions in the gestures, and
6. number checking position of the phone.

The acceleration features are sensitive to large differences in gestural dynamics, while temporal variations are well handled by HMM models. We conclude that facts 1, 5 and 6 are the main factors for the erroneous gesture classifications of testee P2.

4. SUMMARY

Experiments with acceleration sensors were described for static and dynamic gesture recognition of mobile phones. Static gestures were recognized with SOMs while HMMs are used for recognizing dynamic gestures. The results indicate that acceleration sensors provide very useful information for automatic determination of the gestures of the user handling a mobile phone. This information can be used with any portable device that needs to know what is currently happening to it. Possible applications include embedded automatic anti-theft alarm systems, self-monitoring of moving robots, and context-aware embedded digital assistants.

5. REFERENCES

[1] Baum, L. E., Petrie, T., Soules G. and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics, vol. 41, pp. 164-171, 1970.

[2] Boehm, K., Broll, W., Sokolevicz, M., Dynamic gesture recognition using neural networks; a fundament for advanced interaction construction. Proceedings of the SPIE , The international Society for Optical engineering, Vol.2177, p336-46, 1995.

[3] Harrington, M. E., Daniel, R. W. and Kyberd, P. J. A measurement system for the recognition of arm gestures using accelerometers. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, Vol. 209, No.2, pp. 129-133, 1995.

[4] Janet, A. J., Gutierrez R., Chase T. A., White M. W., Sutton J. C., Autonomous mobile robot global self –localization using Kohonen and region-feature neural networks, Journal of Robotic Systems 14(4), pp. 263-282, 1999.

[5] Joutsiniemi S.-L., Kaski S., Larsen T.A., Self-organizing map in recognition of topographic patterns of EEG spectra. Biomedical Engineering, IEEE Transactions on Vol.42, No. 11 , pp. 1062 –1068, 1995.

[6] Kohonen, T. Self-organizing maps. Springer series of information sciences 1st ed. 1995.

[7] Kobayashi, T. and Haruyama, S. Partly-hidden Markov model and its application to gesture recognition. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Vol. 4, pp.3081-3084, 1997.

[8] Lee, C. and Xu, Y. Online, interactive learning of gestures for human/robot interfaces. Proceedings of the IEEE International Conference on Robotics and Automation, Minneapolis, Minnesota, 1996.

[9] Levinson, S. E., Rabiner, L. R. and Sondhi, M. M. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition, Bell System Technical Journal, vol. 62, no. 62, pp. 1035-1074, Apr. 1983.

[10] Levinson, S. E., Rabiner, L. R. and Sondhi, M. M. On the application of vector quantization and hidden Markov models to speaker-independent , isolated word recognition. Bell System Technical Journal, vol. 62, no. 62, pp. 1075-1105, Apr. 1983.

[11] McKenna, S. J. and Gong, S. Gesture recognition for visually mediated interaction using probabilistic event trajectories. Proceedings of the Ninth British Machine Vision, pp. 498-507, 1998.

[12] Min, B.-W., Yoon, H.-S., Soh, J., Yang, Y.-M. and Ejima, T. Hand gesture recognition using hidden Markov models. IEEE International Conference on Systems, Man, and Cybernetics, Vol. 5, pp. 4232-4235, 1997.

[13] Pavlovic, V. I., Sharma, R. and Huang, T. S. Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, July 1997.

[14] Peltola, J. HMM-perustainen puheentunnistus. Diploma thesis, in Finnish. University of Oulu, pp. 48-49, 1998.

[15] Peltola J., Plomp J., Seppänen T. A dictionary-adaptive speech driven user interface for distributed multimedia platform. Accepted to Euromicro workshop on multimedia and telecommunications 1999, Milan, Italy, September 8-10, 1999.

[16] Rabiner, L. R., Juang, B.-H., Fundamentals of Speech Recognition. New Jersey: Prentice Hall, 1993.

[17] Risk M. R., Sobh J. F., Saul J. P., Beat detection and classification EGC using self organizing maps, Engineering in Medicine and Biology Society, Proceedings of the 19th Annual International Conference of the IEEE Vol.1, pp. 89 –91, 1997.

[18] Sawada, H. and Hashimoto, S. Gesture recognition using an acceleration sensor and its application to musical performance control. Electronics and Communications in Japan, Part 3, Vol. 80, No. 5, 1997.

[19] Tuulari, E. Context aware hand-held devices. Espoo: Technical Research Centre of Finland, VTT Publications 412, ISBN 951-38-5563-5, 2000.

[20] Usa, S. and Mochida, Y. A conducting recognition system on the model of musicians' process. J. Acoust. Soc. Jpn. (E), Vol. 19, No. 4, 1998.

[21] Yang, J., Xu, Y. and Chen C. S. Gesture interface: Modeling and learning. Proceedings of the IEEE International Conference on Robotics and Automation, Vol. 2, pp.1747-1752, 1994.