Feature Selection and Multivariate Gaussian Probability Distribution for User Behavior Recognition

Zhiping Zhou, Lele Liu, Xinning Li

Engineering Research Center of Internet of Things Technology Applications Ministry of Education, JiangNan University，Wuxi Jiangsu 214122 China

E-mail: zzp@jiangnan.edu.cn, 6141918006@vip.jiangnan.edu.cn, 18861825208@163.com

Abstract: User behavior recognition using sensory data has become an active field of research in the domain of pervasive and mobile computing. The Principal Component Analysis (PCA) is a common method for feature selection. To obtain the best description and the best classification characteristics of different behaviors, an algorithm of Principal Component Analysis based on Regularized Mutual Information (RMIPCA) is presented. The new algorithm introduces the category information, and uses the sum of regularized mutual information matrices between features under different behavior to replace the covariance matrix. Furthermore, the extracted feature is calculated based on multivariate Gaussian probability distribution, and the transitional noise behavior data is removed according to the probability value. The simulation results show that the performance of the proposed is better than the others compared, which can effectively identify the user's daily behavior.

Key Words: Mutual Information, inter-class information, Multivariate Gaussian Probability Distribution

1 Introduction

In recent years, with the progress of information science and sensor technology, human behavior recognition based on sensors has obtained a great development, in which the wearable sensors has a wide broad application prospect[1]. Behavior of users contains rich context situation information, and in a short period of time, which can reflect the current active state and intention of the individual. After a long time, the characteristic of users' behavior can reflect different individual's favor and habit[2]. Therefore, the perception of user behavior is to provide an important basis for personalized service. Kwapisz et.al[3] used 29 individual data for unified modeling, which could identify the daily activities such as walking, running, going up and down. Altun et al.[4] put five Inertial Measurement Units (IMU) set up on pedestrians to identify different behaviors. The work in [5-8] used fixed length of sliding window to segment data, and then identify the behavior. However, the method in [5-8] does not consider the noise behavior data during the transition period between the behaviors when using the sliding window to segment the data, so it has a certain influence on the behavior recognition accuracy.

The extracted feature vector usually exhibit high dimensional properties in user behavior recognition. The high dimensional data not only leads to increase the computational complexity, but also inconducive to reduce the redundant information in the feature space. Wang et al.[9] collected human behavior data by the wearable sensor. Nineteen activities were classified, including lying, sitting, standing, walking, going upstairs, downstairs, etc. The Principal Component Analysis (PCA) method was adopted to reduce the extracted 630-dimensional feature vector to 42 dimensions, which reduced redundant information in the feature space. In [10], the PCA method was utilized to select subset of the extracted feature, and then the motion and traffic pattern activities were identified including stationary, walking, riding and cycling. However, the method in work [9-10] can only measure the linear relationship between features when PCA is utilized for feature selection. In the process of information compression, the intra-class change information of obvious difference may be preserved, but the inter-class information may be abandoned due to the low difference, which affects classification precision.

In this work, we introduce the regularized mutual information to measure the interdependence between the features, which could provide more linear and non-linear relation information between features. With the supervisory information added to the feature selection, a new matrix is constructed by using the mutual information matrix of the features in each category. Furthermore, the transitional noise behavior data are removed according to the probability value of the feature based on the multivariate Gaussian probability distribution, which could solve the influence of the transition noise behavior data on recognition accuracy.

2 Gesture segmentation and feature extraction

2.1 Gesture data segmentation

In the process of data collection, due to the hardware imperfections during the sensor manufacturing process, every sensor has power noise and random dithering. Besides, the noises generated by different sensor embedded in different mobile devices are different in amplitude. The moving average filter[11] is utilized to preprocess the original data, which can achieve better smoothing effect and reduce the noise interference. In this work, the gesture segmentation begins with the user touching the screen, when the user operates the mobile phone screen, the acceleration sensor and the touch screen information will change obviously. In the four behavior of browsing web (BW), chatting type (CT),

*This work is supported by the Special Funds of Basic Research Business Expenses of Central University under Grant No.JUSRP51510.

picking up (PU) and playing games (PG), the users makes a call by sliding the start button and click the end button to hang up the phone. Statistical analysis was performed on the collected sample data, the time interval before and after touching the screen is more than 10s in process of making calls, therefore, we select the time interval greater than 10s to get gesture signal interception, which is shown in Fig 1. The intercept signal is divided into block1, block2, where block1 including 3 kinds of behavior, namely browsing web pages, chatting and playing games. The block2 is the behavior of making call. In the process of block1, transition phase stands for the two kinds of operation behavior of transition time.

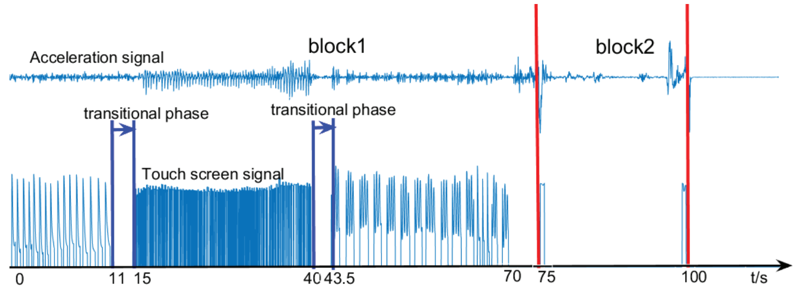

Fig 1. Interceping gesture signal

2.2 Feature extraction

In order to make the gesture signal can effectively identify the behavior, it is necessary to extract feature from the acceleration and touch screen signal. The features of the extracted acceleration signal include four kinds, which divided into time domain and frequency domain characteristics, and the touch screen signals were extracted from seven kinds of features. The a_x , a_y and a_z is the acceleration signal on X, Y, and Z axis, and x, y is touch screen coordinate sequence. The following features are extracted from the acceleration and the touch screen signal in each sliding window with the length l. The work is set to the fixed sliding window size of 10s with 50% overlapping.

(1) Gesture Energy E: The intensity of the gesture during the motion is called the gestural energy, which is calculated as the sum of all three-axes accelerations from beginning to the end of the gesture.

(2) Peak Value K: The peak value reflects the steepness of the acceleration sensor signal at the crest, and the peak value of a_x as shown in formula (1):

$$K = \frac{\sum_{i=1}^{l}(a_{xi}-\mu)^4 f}{\sigma^4} \tag{1}$$

where μ is the mean of the acceleration, σ the variance of the acceleration, and f_i is the interval of sliding window.

(3) The acceleration peak number P: the number of acceleration peaks is the sum number of three-axis acceleration peak. The number of X axis peak is shown in Fig 2, and the Y, Z axis is the same.

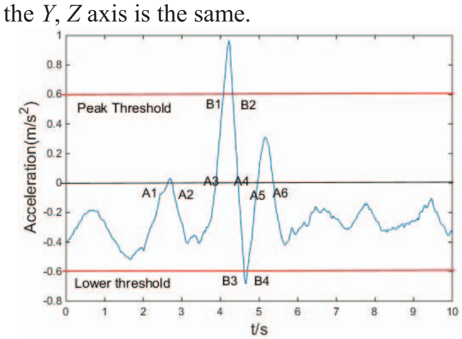

Fig 2. Number of acceleration peaks

First, we find out all the intersection points of the acceleration data with the X-axis. Then the points B_1, B_2, B_3 and B_4 are found between these two adjacent intersections. And the number of peaks above the threshold is the acceleration peaks.

(4) FFT coefficient FT: The FFT coefficient is a typical frequency domain feature of the acceleration signal, which reflects the frequency component amplitude of the signal. For a data window of length l, the a_x Fourier transform is:

$$FT = \sum_{i=0}^{l-1} a_{xi} e^{-j\frac{2\pi}{l}ik},\ k = 0,1,\ldots,l\text{-}1 \tag{2}$$

The first 32-dimensional coefficients of acceleration we need in this work is shown in Fig 3.

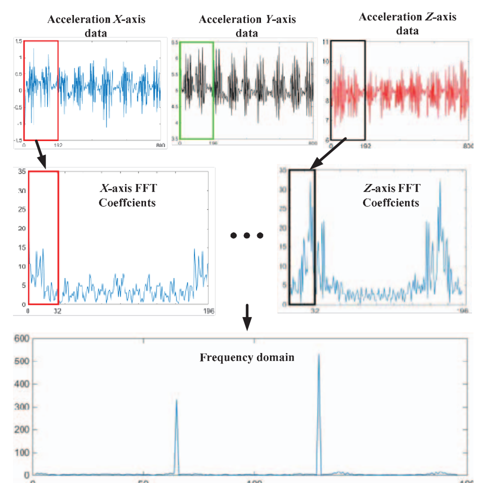

Fig 3. FFT coefficients of the acceleration signal

(5) Sliding speed S: The sliding speed reflects the moving speed of the finger on the screen when the mobile phone user operates different programs, as shown in formula (3):

$$S_i = \frac{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}}{(t_i - t_{i-1}) * \theta_s}, i \in 2,3,\ldots,l \tag{3}$$

$$\theta_s = \sqrt{(x_{\max} - x_{\min})^2 + (y_{\max} - y_{\min})^2} \tag{4}$$

(6) Slide direction ∂ : The movement direction of the gesture trajectory at point (x_i, y_i) is defined as:

$$\partial = \arctan \frac{y_i - y_{i-1}}{x_i - x_{i-1}} \tag{5}$$

(7) Gesture sliding length S_L:

$$SL = \sqrt{\frac{(x_{end} - x_{start})^2 + (y_{end} - y_{start})^2}{(x_{\max} - x_{\min})^2 + (y_{\max} - y_{\min})^2}} \tag{6}$$

(8) Slope of the user's sliding gesture S_C:

$$SC = \arctan(\frac{y_{end} - y_{start}}{x_{end} - x_{start}}) \tag{7}$$

In addition, the mean value S_{mean} and the variance S_{std} of the sliding velocity are extracted, as well as the number S_n of the slide gesture. In this way, we can get the feature vector $F = [E, K, P, FT, S, \partial, S_L, S_C, S_{mean}, S_{std}, S_n]$ of each sliding window, which can be combined into a feature data set $F \in R^{M \times N}$, where M is the number of samples, and N is the dimension of each feature vector.

3 Proposed Method

Considering the imbalance of the feature distribution, the extracted acceleration signal and the touch screen signal feature come from the time domain and frequency domain, and their values are quite different, the z-score[12] is used to normalize the exacted feature vector in this work. However, the normalized feature vector set still has some problems such as high dimensionality and large redundant information. In order to optimize the feature space and reduce the invalid information components, PCA is utilized to select the feature subset.

3.1 Principle Component Analysis

As an unsupervised dimensionality reduction method, PCA is an effective and widely used method in behavior recognition system. It is based on the covariance matrix, by calculating the correlation coefficient between each feature. However, PCA can only measure the linear relationship between the features and ignores the class labels that are essential to retain the discriminative information for the subsequent classification[13].

Mutual information can measure the degree of interrelatedness between features, which indicates the amount of information that is shared between features, and mutual information can be used to evaluate the linear and non-linear relationship between features. In this work, the mutual information is introduced to measure the interdependence of features, to provide more linear and non-linear relationship information between features. Meanwhile, the supervising information is added in the feature selection, the covariance matrix of the feature is replaced by the sum of the mutual information of the features in each category condition. Then the new method is applied to the user behavior recognition.

3.2 The RMIPCA Method

Mutual information (MI) measures the degree of interrelationship among the features in $F \in R^{M \times N}$, and indicates the content of common information between features. The value set for the feature $F_i \mid f_i^u$, $u = 1, 2, \cdots, M$ is $\{f_i^1, f_i^2, \cdots, f_i^{M_I}\}$, which represents the M_I subset. The mutual information $\mathrm{MI}(F_i, F_j)$ is defined as (8):

$$\mathrm{MI}(F_i, F_j) = \sum_{m=1}^{M_I} \sum_{n=1}^{M_J} p(f_i^m, f_j^n) \log \frac{p(f_i^m, f_j^n)}{p(f_i^m) p(f_j^n)} \tag{8}$$

The mutual information is 0, when F_i and F_j completely independent each other, which means that there is no similar. On the contrary, the higher degree of correlation between feature F_i and F_j is, the greater value of mutual information is. Therefore, the correlation between the characteristics of the information can be well measured by using mutual information.

Regularized Mutual Information (RMI) is the normalization of mutual information. Standardized mutual information compensates the deviation of mutual information to multi-valued variables, and the standardized information value is strictly limited to interval [0,1]. Regularized mutual information is defined as:

$$\mathrm{RMI}(F_i, F_j) = 2R = 2\frac{\mathrm{MI}(F_i, F_j)}{H(F_i) + H(F_j)} \tag{9}$$

where $H(F_i)$ is the entropy of feature F_i, $H(F_i)$ is the entropy of feature F_j.

A new feature selection algorithm of PCA based on regularized mutual information (RMIPCA) is proposed in this paper. The correlation coefficient matrix is replaced by regularized mutual information, and the supervising information is added for better application to classification. A new matrix is constructed by summing the regularized mutual information matrices between features in each category, as shown in formula (10):

$$D = \sum_{l=1}^{c} \sum_{F_i, F_j \in l} \mathrm{RMI}(F_i, F_j) \tag{10}$$

where c indicates the total number of categories, D is the regularized mutual information matrix. The diagonal elements represent the self-information of the feature, that is, the information entropy, and the non-diagonal elements represent the mutual information between the features. Whether the mutual information or information entropy are real, when there is no correlation between the two characteristics, the value of mutual information is 0, otherwise positive, thus D is a non-negative real number matrix. Meanwhile, mutual information satisfies symmetry, so D is a non-negative real symmetric matrix.

Similar to the PCA, the problem of solving the projection vector of RMIPCA can be converted to the following eigenvalue problem, as shown in formula (11):

$$D\alpha = \gamma\alpha \tag{11}$$

where the eigenvector $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_k]$ corresponding to the top k largest eigenvalue $\gamma_1 \geq \gamma_2 \geq \cdots \gamma_k$ is the projection vector of the RMIPCA. The k dimension principal component contribution rate μ_k of the RMIPCA is defined as the ratio of the principal component to the total principal component information, and the cumulative contribution rate of the k-th principal component is the sum of the first k principal component contribution, as shown in the formula (12):

$$\mu_k = \frac{\gamma_k}{\sum_{k=1}^{N} \gamma_k}, \ \xi_d = \sum_{i=1}^{k} \mu_k \tag{12}$$

where N is the original dimension of the dataset, the former d principal component with ξ_d of 85% ~ 95% is selected as the new feature, which is expressed as $\chi = \{x_1, x_2, \cdots, x_d\}^T$.

4 Multivariate Gaussian Probability Distribution

The sliding window is utilized to segment the gesture data, and the gesture signals which have been segmented may contain two kinds of transition noise data of behavior. The noise data mentioned before is hidden in the middle of the data segment, as shown in Fig 4. In order to effectively remove the influence of transitional data on behavior recognition, a Multivariate Gaussian Probability Distribution (MGPD) method is utilized to calculate the feature after dimension reduction and the maximum posterior probability is used to judge the classification. Further, according to the probability value of the feature we can remove the transitional noise behavior.

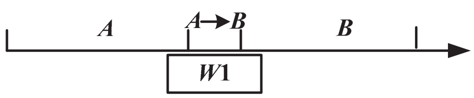

Fig 4. The transitional behavior

The core of the multivariate Gaussian probability distribution method is based on the Bayes rule:

$$p(A_j \mid \chi) = \frac{p(\chi \mid A_j)p(A_j)}{p(\chi)} \tag{13}$$

where $\chi = \{x_1, x_2, \cdots, x_d\}^T$ is the d dimension vector after dimension reduction, $p(A_j)$ is the prior probability of the j-th behavior state. In this work, the prior probabilities $p(A_j)$ of all kinds of behaviors in the training samples are equal. The $p(\chi \mid A_j)$ is the likelihood function, which expresses the probability of the χ feature under the behavior A_j. The $p(A_j \mid \chi)$ indicates the conditional probability that the sample χ belongs to behavior A_j, which called posterior probability. By the formula (13) shows that the likelihood function $p(\chi \mid A_j)$, which can compare the probability of d-dimensional feature χ belongs to the different behavior A_j.

The acceleration sensor signal and the touch screen signal is modeled as Gaussian distribution[14], and the probability of the extracted d-dimensional feature $\chi = \{x\}$ calculated by using the multivariate Gaussian distribution, can be expressed as:

$$\begin{aligned} p(\chi \mid A_j) &\propto p(\chi; \mu_j; \textstyle\sum_j) \\ &= \frac{1}{(2\pi)^{d/2} \mid \sum_j \mid} e^{-\frac{1}{2}(\chi-\mu_j)^T \sum^{-1}(\chi-\mu_j)} \end{aligned} \tag{14}$$

where μ_j is the d dimension mean matrix and Σ_j is the $d \times d$ dimension covariance matrix.

The feature of signal is extracted by the sliding window, and the RMIPCA is utilized for feature selection. Furthermore, the MGPD is utilized to identify the behavior, and the threshold value is set to remove the transitional noise behavior data according to the probability value of the feature in the class condition, as shown in Fig 5. The sliding window W1 extracts the signal characteristic χ_{W1}, and the probability of determining the behavior A by MGPD is $p_{\mathrm{W1}}(\chi \mid A) = 2.52$. Similarly, the probability that the features sliding window W2 and W3 extracted belong to behavior A respectively is $p_{\mathrm{W2}}(\chi \mid A) = 1.66$ and $p_{\mathrm{W3}}(\chi \mid A) = 0.002$. When the probability value of the feature in the class condition satisfies $p_{\mathrm{W}i}(\chi \mid A) < \delta_{th}$, it is determined as the transition noise behavior data, and the tested δ_{th} is equal to 0.006 when the threshold is set to remove the transition noise behavior data.

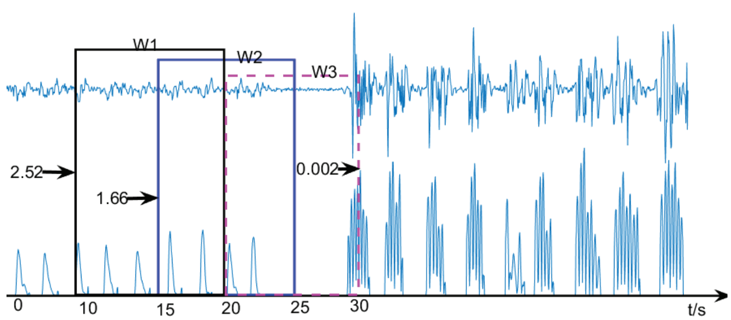

Fig 5. Removing transitional behavior data

The steps of proposed method can be summarized as follows:

$\Delta t_{x_i, x_{i+1}}$ indicates the time interval between the first touch and the second touch

1. if $\Delta t_{x_i, x_{i+1}} \geq 10s$ then
2. Segmenting signals according to the touch screen
3. else
4. Extracting the gesture signal features
5. Feature selection by RMIPCA
6. Judgment of behavior by MGPD
7. Calculating probability density $p_{\mathrm{W}i}(\chi \mid A)$
8. if $p_{\mathrm{W}i}(\chi \mid A) < \delta_{th}$
9. Determining the transition to the two behaviors
10. Removing the transient behavior noise data

11. end if
12. end if

5 Experimental Results and Discussion

The subsequent simulation was based on MatlabR2014a in a PC computer with AMD A4-3305M 1.90GHz CPU and 2.00G memory. Fifteen experimenters (10 males and 5 females) are selected to collect the mobile phone sensor data records, which provide data and labels for experiment, and obtain the sensor record data for 7 consecutive days. The sampling frequency of the acceleration sensor is set to 50Hz, the touch screen sensor is collected at 0.02s time, which maintains the same frequency with acceleration sensor. If the mobile phone touch screen without gestures, x and y axis is stored as 0. A total of 5,600 valid gesture data were collected, and each behavior has 1400 samples.

5.1 Determination of Optimum Dimension

In this work, we compare the effectiveness of the proposed RMIPCA with two representative feature selection algorithms which are PCA[9] and LDA[15]. LDA is a global supervised algorithm which is wildly used in the stage of data preprocessing. For the 4 kinds of behavior, the behavioral feature vector has 169 dimensions. The MGPD method is used for recognition, and the PCA, LDA and RMIPCA on the accuracy rate changes with the situation, as shown in Fig. 6.

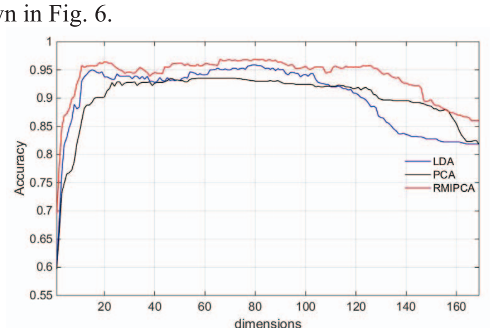

Fig 6. Recognition accuracy of the algorithms

It is observed that with the increase of the dimension, the recognition accuracy trend of the algorithms increases first, then tends to be stable, and finally decreases. The RMIPCA can achieve a relatively stable state in 23 dimensions. At this time, the cumulative contribution rate of RMIPCA is over 85%, and the recognition accuracy is 96.52%. In order to show the superiority of the proposed algorithm in comparison with the other two algorithms in terms of effectiveness and real-time, we choose 23 dimensions as the optimal dimension, and the other two algorithms get the corresponding dimension when the optimal accuracy is reached. Therefore, the PCA is selected as 57 dimensions and the LDA is selected as 79 dimensions to achieve the best accuracy, and the accuracy rate is 94.26%, 95.57%.

5.2 Result analysis

In this section, we explain the experimental results we reached during the experiments. We used 10-fold cross-validation for the experiments and applied accuracy parameter for evaluation of the methods. The average accuracy rate of 10 times is taken as the accuracy of the final recognition behavior. We compared the methods such as PCA, LDA and the method in literature [8] without optimizing processing for the extracted features. The *recall*, *precision*, *accuracy*, and *F-score* values of the four algorithms are shown in Table 1.

Table1. Comparison Results (%)

Behavior	Index	[8]	PCA	LDA	RMIPCA
BW	*recall*	84.82	0.9175	93.18	94.21
	precision	87.09	0.9210	93.74	95.30
	F-score	85.94	0.9192	93.46	94.75
PU	*recall*	100	100	100	100
	precision	98.66	99.43	99.50	99.57
	F-score	99.32	99.71	99.75	99.78
CT	*recall*	90.57	93.07	95.34	96.50
	precision	91.41	95.45	95.63	96.85
	F-score	90.98	94.24	95.48	96.67
PG	*recall*	88.43	92.41	93.86	95.36
	precision	86.38	90.17	93.39	94.35
	F-score	87.39	91.28	93.62	94.85
Accuracy	*AC*	90.91	94.26	95.57	96.52

The literature [8] without optimizing processing for the extracted features which has 169 dimensions, the number of correctly and incorrectly classified feature vectors with 10-fold cross validation is given in Table 1. The overall correct classification rate of this method is 90.91%. The feature vector in literature [8] contains many redundant and irrelevant data, which not only leads to increase the computational complexity, but also reduce the classification accuracy. The PCA could preserve the high ability of characterization in original data, and the feature has the best description ability. The overall correct classification rate of PCA is 94.26%, which is higher than that in literature [8]. In the process of feature selection using LDA, the information of inter-class is considered. Compared with PCA, LDA significantly improves the performance of the classification. Although the LDA consider inter-class change information for the feature selection, the RMIPCA introduces mutual information to measure the interdependence between the two features, and providing more linear and nonlinear relation information among the features, the recognition accuracy is up to 96.52%.

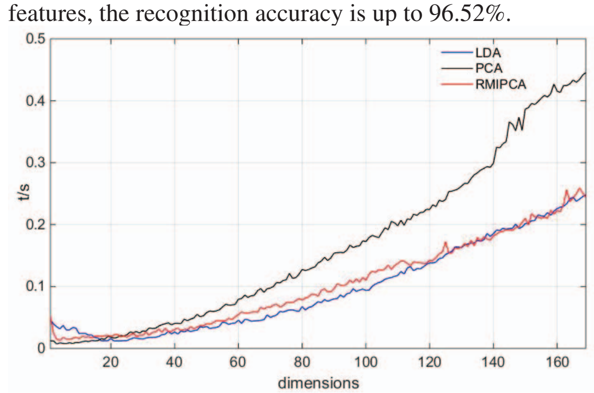

Fig. 7. Training time in different dimensions

According to Fig. 7, it is observed that with the increase of dimension, the training time will be increased. The training time of RMIPCA is less than PCA, and is basically the same as LDA. Combining with RMIPCA and other two algorithms to identify performance results, the proposed method effectively improves the recognition accuracy and shortens the training time, which reflects the real-time and validity in user behavior classification task.

6 Conclusion

In this work, the acceleration sensor and touch screen data is utilized to analyze the behavior of users in browsing web, chatting, making calls and playing games. By introducing the regularized mutual information into feature selection, the covariance matrix of the feature is replaced by the sum of the regularized mutual information between the features in each class condition. The posterior probability of the reduced feature is calculated by using the multivariate Gaussian probability distribution method, and the maximum probability is to judge the classification. The transitional noise behavior is further removed according to the probability value of the feature. Experimental results show that the proposed method can effectively improve the correct rate of behavior recognition.

References

[1] Hemminki S, Nurmi P, Tarkoma S. Accelerometer-based transportation mode detection on smartphones, *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, 2013: 13.

[2] L. Chen, S. Li, G. Pan. Smartphone: Pervasive Sensing and Applications, *Chinese Journal of Computers*, 2015,02:423-438.

[3] Kwapisz J R, Weiss G M, Moore S A. Activity recognition using cell phone accelerometers, *ACM SigKDD Explorations Newsletter*, 2011, 12(2): 74-82.

[4] Altun K, Barshan B, Tunçel O. Comparative study on classifying human activities with miniature inertial and magnetic sensors, *Pattern Recognition*, 2010, 43(10): 3605-3620.

[5] Chiang J H, Yang P C, Tu H. Pattern analysis in daily physical activity data for personal health management, *Pervasive and Mobile Computing*, 2014, 13: 13-25.

[6] Adaskevicius R. Method for Recognition of the Physical Activity of Human Being Using a Wearable Accelerometer, *Elektronika ir Elektrotechnika*, 2014, 20(5): 127-131.

[7] Hu L, Chen Y, Wang S, et al. b-COELM: A fast, lightweight and accurate activity recognition model for mini-wearable devices, *Pervasive and Mobile Computing*, 2014, 15: 200-214.

[8] Catal C, Tufekci S, Pirmit E, et al. On the use of ensemble of classifiers for accelerometer-based activity recognition, *Applied Soft Computing*, 2015, 37: 1018-1022.

[9] Wang L. Recognition of Human Activities Using Continuous Autoencoders with Wearable Sensors, *Sensors*, 2016, 16(2): 189.

[10] Bhattacharya S, Nurmi P, Hammerla N, et al. Using unlabeled data in a sparse-coding framework for human activity recognition, *Pervasive and Mobile Computing*, 2014, 15: 242-262.

[11] Khan A M, Lee Y K, Lee S Y, et al. Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis, *IEEE International Conference on Future Information Technology*, 2010: 1-6.

[12] Han J, Pei J, Kamber M. Data mining: concepts and techniques, *Elsevier*, 2011.

[13] Tao D, Jin L, Yuan Y, et al. Ensemble manifold rank preserving for acceleration-based human activity recognition, *IEEE transactions on neural networks and learning systems*, 2016, 27(6): 1392-1404.

[14] Yu N, Zhao Y, Han Q, et al. Identification of partitions in a homogeneous activity group using mobile devices, *Mobile Information Systems*, 2016, 2016:1-14.

[15] Dufrenois F, Noyer J C. Formulating robust linear regression estimation as a one-class LDA criterion: Discriminative hat matrix, *IEEE transactions on neural networks and learning systems*, 2013, 24(2): 262-273.