Comput Syst Sci & Eng (2017) 5: 383–396


International Journal of
Computer Systems
Science & Engineering

# A context-aware hierarchical approach for activity recognition based on mobile devices

**Shugang Zhang**[1] **, Zhiqiang Wei**[1]**, Jie Nie**[2]**, Lei Huang**[1] **and Zhen Li**[1,*]

[1] *College of Information Science and Engineering, Ocean University of China, Qingdao, China.*
*E-Mails: zhangshugang@hotmail.com; weizhiqiang@ouc.edu.cn; ithuanglei@gmail.com; niejie@tsinghua.edu.cn*
[2] *Department of Computer Science and Technology, Tsinghua University, Beijing, China*

Various sensors integrated in the wearable device provide massive data for activity recognition. In this paper, a context-aware hierarchical approach is proposed for the recognition of activities using accelerometers on smartphones and smartwatches. We adopt a simple variance threshold based method and separate the activities into two major categories named body-fixed set and body-unfixed set according to the inherent characteristics of these activities in the first layer. Next, the Support Vector Machine approach is used respectively for the two sets in the second layer. A probability distribution over activity labels instead of a single activity result is generated in this layer. In the third layer, the contextual information is introduced to improve the classification result. Our comparative study with ordinary Support Vector Machines and other alternative methods has shown that our method is more robust and accurate.

Keywords: Activity recognition; Smartphone; Smartwatch; Support Vector Machine; Accelerometer

## 1. INTRODUCTION

Human activity recognition (HAR) aims to recognize activities from a series of observations on the actions of subjects and the environmental conditions. It is the basis of many areas such as health care [1–3], Smart Home [4–7], Human-Computer Interaction (HCI) [8], and ubiquitous computing [9]. Lots of work has been done on HAR using accelerometers [9,10]. Some of these methods focused on using multiple accelerometers [11–13], other methods attempted to utilize only one single accelerometer [14–17] and determine the optimal placement of it [18]. With the popularization of smartphones, separated accelerometers are replaced by sensors integrated in the device. These smart devices provide a more convenient way to collect data.

Multi-sensor performs better than single-sensor [11,19], since these sensors can capture data from different places at the same time. However, subjects may feel obtrusive using current multiple accelerometers capture systems since there are so many devices attached on the body [20]. On the other hand, from a single accelerometer or unobtrusive smartphone, most studies could only recognize several simple activities such as walking, standing due to the limitation of position and the number of sensors. Not many methods have tried to combine additional wearable devices with smartphones [21,22].

Due to the limited information that accelerometers provide, especially the single accelerometer method, some methods try to recognize complex activities by utilizing environmental attributes [12,20,23]. These attributes include temperature, humidity, audio level, and the location information obtained by GPS. Among these attributes, GPS data is important part of the contextual information, and they are widely used in the context-aware activity recognition. However, GPS sensors perform far from being satisfactory indoors which affects the performance of HAR.

In this paper, a hierarchical method is proposed to optimize the activity classifier by dividing the process of recognition into three stages. We utilize two mobile devices (i.e. smartphone and

*Corresponding author. E-mail: lizhen0130@gmail.com; Tel.: +86-532-66781712; Fax: +86-532-66781687

smartwatch) and various embedded sensors for recognition. The pipeline of the method is shown in Figure 1. First in the *Body Intensity (BI) layer*, the hierarchical classifier divides the activity dataset into two sets, namely body-fixed set and body-unfixed set, according to the motion of the thigh. The term "body-fixed" means the torso of subject is in a "fixed" state or show slight fluctuation and sometimes with limb movements (e.g. reading, stretching), while the "body-unfixed" indicates an "unfixed" and moving torso (e.g. running, ascending stairs). This layer utilizes the inherent characteristics of two major categories of activities and these characteristics could be captured by our data collection scheme. Then in the *Support Vector Machine Classification (SC) layer*, the SVM is utilized to classify the two sets respectively. A probability distribution over activity labels instead of the final activity result is generated in this layer, preparing for the further contextual information fusion in next stage. Finally in the *Contextual Fusion (CF) layer*, we propose a novel systematic approach which combines the activity contextual information with the probability distribution in last layer to recalculate the probability and improve the accuracy. A naÃ¯ve Bayes model is presented for the fusion of time and location information. In order to deal with the GPS signal missing problem when subject performs activities indoors, a WiFi-assisted GPS labeling method is proposed to acquire the location information of indoor activities.

The rest of the paper is organized as follows. We discuss the related work in section 2. Section 3 describes the extracted features used in this paper, as well as their extraction methods. Then the context-aware hierarchical approach is presented in section 4. Three layers are demonstrated respectively in this section. Section 5 presents a series of comparative evaluations to validate the effectiveness of the proposed approach. Finally, we conclude the paper in section 6.

## 2. RELATED WORK

Accelerometer-based activity recognition has been extensively investigated. The majority of the past work mainly focused on three goals. The first and most important goal is to recognize more activities and achieve high recognition accuracy. This is usually addressed by placing multiple sensors on subjects. However, subjects may feel obtrusive wearing too much sensors across the whole body, which is the second problem that needs to be addressed especially when the implemented recognition system is actually applied in real life. The third focus of past work is the context utilization. Rational utilization of contextual information can be an effective way to improve accuracy. In this section, related works and solutions concerning the mentioned three goals are discussed.

To enhance the recognition performance and recognize more complex activities, previous work tends to use multiple accelerometers or additional sensors. Multiple accelerometers prove to be more accurate comparing the single accelerometer solution when recognizing complex activities, since accelerometers placed at different body positions provide richer limb movement data and capture more details. For example, the sensor on the dominant wrist could provide critical information for some daily activities such as *brushing teeth* or *waving hands* [11,22,24]. Bao & Intille [11] used five sensors distributed over the whole body. Through C4.5 Decision Tree classifiers, they recognized up to twenty activities with overall recognition rate 84.26%. Tapia et al. [22] used five tri-axial wireless accelerometers and a wireless heart rate monitor, and the recognition accuracy of their method is 94.6% on their subject-dependent gymnasium activity sets also using C4.5 classifiers.

Recent deep learning methods have made breakthroughs in image and speech recognition. For multi-accelerometer based activity recognition, several pioneering works have reported state-of-the-art performance on several benchmark problems using deep convolutional neural networks (convnet) and their automatic feature extraction mechanism. Zeng et al. [25] is the first one who use convnet and automatic feature extraction for activity recognition. One convolutional layer structure and partial sharing weight is used and achieves good results on several benchmark datasets. Yang et al. [26] constructed a deep convnet with three convolutional layers and further improve the recognition performance. In [27], a deep framework for activity recognition based on convolutional and LSTM recurrent neural networks is proposed, which shows apparent advantage of distinguishing similar kind activities, as well as the ability of fusing homogeneous sensor modalities. This research further improves the F1 scores in some benchmark datasets.

Although better results and more activity types can be obtained by adding more accelerometers or other sensors especially combining the recent deep learning approaches, the number of accelerometers is, in fact, completely ignored. Even up to 19 sensors were used across the body when collecting data [28].

Obviously, subjects may feel obtrusive wearing too much sensors across the whole body. To alleviate such uncomfortable feeling and reduce the redundant sensors, the optimal placement of accelerometer is discussed and most work consider the wrist and thigh (trouser pocket) as the best positions. Bao & Intille [11] found that just using the accelerometers on the thigh and wrist did not decrease the recognition performance apparently. In [29] a dynamic Bayesian networks for the exemplary application activity recognition is proposed, which compared the performance of accelerometers for different parts of the body and decided the belt, or waist, as the best place. In [30] an autoregressive (AR) model is utilized to recognize four activities, and it found that the best place is the trouser pocket. Other studies suggested that the optimal placement is wrist [22,31]. In fact, the optimal placement depends on the type of activity. For the ambulation activities, the accelerometer on the belt or in the trouser pocket is enough. However, the same location could not provide sufficient information in recognizing activities involving hands.

Another way of dealing with obtrusiveness is using accelerometers embedded in wearable devices. These devices are usually necessary to carry (e.g. smartphone) or used to replace traditional wearing products (e.g. smartwatch) [32]. People are accustomed to wearing these devices comparing to multiple sensors, therefore these mobile devices provide an unobtrusive way to capture data. Moreover, people would like to put their phones in the trouser pocket which is one of the optimal placements of accelerometer mentioned before. In [33] an activity classifying system was built relying on sensors within a single smartphone. Low-cost and lightweight classifiers were evaluated and optimized to achieve a better result in the constrained computational

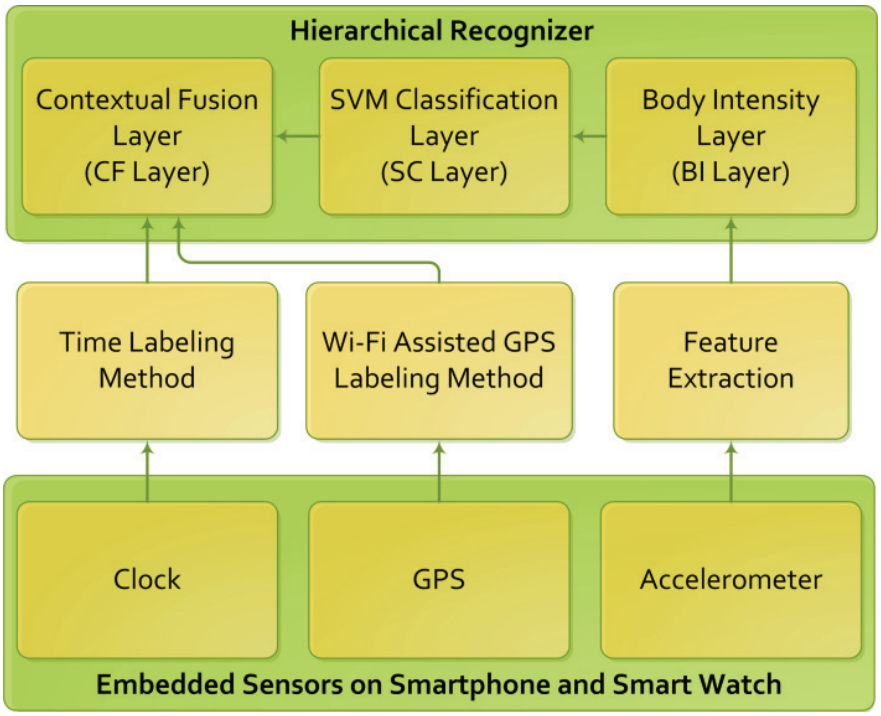


**Figure 1** Overall architecture of the proposed method.

resources. Kwapisz et al. [34] used a smartphone to recognize five activities including standing, walking, jogging, ascending stairs, and descending stairs. Three classification techniques including J48 Decision Trees, Logistic Regression and multilayer Neural Networks were used and achieved high levels of accuracy in most cases. Khan et al. [35] proposed a smartphone-based recognition method which using Kernel Discriminant Analysis to address the high within-class varies problem that the accelerometer varies for the same activity due to the random placement of phone. Five activities were recognized with high average accuracy through Artificial Neural Network. Ronao and Cho [36] successfully applied the recent deep learning approach on smartphone-based activity recognition. A multi-layer convnet is proposed to automatically extract features from raw time-series sensor data. Cumbersome feature hand-crafting is omitted. The system was evaluated on a dataset composed of six activities and achieved the state-of-the-art performance. Some other methods focus on additional sensors combined with smartphone for activity recognition. In [19] a smartphone accelerometer paired with a dedicated chest sensor is used to detect only six types of activity. Similarly, Lara et al. [37] used an additional strap placed on the chest and also detected only six activities.

It is worth mentioning here that although smartphone-based methods can, to a great extent, address the obtrusiveness, most of these methods can only recognize limited and simple activity categories (normally 4∼6 activities). This leads to a compromise problem between single accelerometer solution and multi-accelerometer solution. Moreover, in spite of the great potential of smartphones or multi-accelerometer network when used for recognition, these "context-missing" approaches often perform badly and in some cases even completely confused to recognize any high level activities such as *brushing teeth* or *eating*. For these reasons, contextual information is considered.

People tend to do certain activities in specific time and locations. Such contextual information is explored to assist in recognizing activities and improving accuracy. Liao et al. [12] extracted activities from traces of GPS data using hierarchical Conditional Random Fields. By generating a consistent model of activities and places of a person, some contextual activities such as working and getting on bus were able to be recognized. In [13] an ontological reasoning method through receiving location information from GPS sensors is introduced for complex activity recognition. Adding location information or labeling significant places is helpful to recognize more activities. However, their method suffers failure indoors as the GPS performs badly indoors because of signal missing. Han et al. [20] proposed a hierarchical activity recognition framework including a location-aware engine. Outdoor activities are enriched due to the introduction of GPS data. All indoor locations without GPS signal are considered to be home or office, and these two places are not distinguished from each other. Only three activities (i.e. walking, sitting and standing) are supposed to happen in these indoor places.

In summary, the work is motivated by the following limitations of activity recognition. On one hand, the multiple accelerometer methods could recognize more activities with a high accuracy, but subjects may feel obtrusive when wearing too much sensors. On the other hand, high accuracy and unobtrusiveness has smartphone or other single accelerometer based methods (suppose the single accelerometer is worn at the optimal placement) achieved, most previous work, however, reported limited number of activities. Motivated by these existing limitations, we propose an approach to recognize 13 daily activities in an unobtrusive way. Using two common mobile devices, smartphone and smartwatch, a compromise is achieved which has high accuracy, unobtrusiveness, and the ability of recognizing multiple

and complex activity categories. Adopting a hierarchical scheme and a novel systematic approach, the contextual information is fused rationally to further improve the recognition accuracy, especially those high level activities.

## 3. FEATURE EXTRACTION

### 3.1 Activities

Thirteen activities were studied, include eating, typing, drinking, waving hands, reading, stretching, brushing teeth, sleeping, washing hands, walking, running, ascending stairs, descending stairs (Table 1). We choose these activities since these thirteen activities could cover common daily activities and represent life style of a person. During the data capture, nearly half of these activities have long duration, including eating, typing, reading, brushing teeth, sleeping, washing hands, walking, and running. The rest activities have a very short duration, thus we required subjects to repeat multiple times.

### 3.2 Feature extraction

In order to extract features from data, all data is regularized into 7-second samples. For long duration activities such as walking and running, data is divided into 7-second segments which are sufficient to capture pattern in these periodic activities. Each segment is used as an example. For the short duration activities such as ascending stairs and drinking, most cases last less than 7 seconds thus the activity process is directly collected within a 7-second window in collection stage. Each 7-second sample is used directly for feature extraction without further dividing. The short duration activities cannot be divided because the features extraction method of them needs at least a complete process, not a fragment of it. Features that extracted from the raw data are presented in Table 2. Collection time and location are also considered as contextual attributes.

Standard statistical metrics mean and variance of each axis (Feature1-4) are used as features because of their good performances. The mean value of each axis can accurately reflect the posture of body or wrist, for example, sitting or standing can be detected by the mean value of the accelerometer data of thigh. The variance of each axis can reflect the current exercise intensity. The energy value (Feature 6, 7) [16] is also introduced and it is calculated through the sum of the squared discrete FFT component magnitudes of the signal, and the sum is normalized by a certain window as presented in (1)

$$energy_x = \frac{\sum_{i=1}^{|x|} |x_i|^2}{w} \tag{1}$$

where $x$ is the i*th* FFT component of the window for $x$ axis and $w$ is the window length.

In addition, the sum of variance and energy values of the phone for three axes (Feature 5, 8) are calculated respectively, as presented in (2) and (3).

$$sum_v = var_x + var_y + var_z \tag{2}$$

$$sum_e = energy_x + energy_y + energy_z \tag{3}$$

where $sum_v$ is the sum of the 3-axis variance and $sum_e$ is the sum of the 3-axis energy. $var_x$, $var_y$, $var_z$ are variance of $x$, $y$ and $z$ axis. $energy_x$, $energy_y$, $energy_z$ are energy of $x$, $y$ and $z$ axis.

The sum is calculated to determine the threshold in BI layer. Finally, the collection time label (Feature 9) and collection location label (Feature 10) are extracted from the raw data as feature of contextual information.

## 4. THE CONTEXT-AWARE HIERARCHICAL APPROACH

There are three levels in the proposed hierarchical approach named Body Intensity (BI) layer, SVM Classification (SC) layer, and Contextual Fusion (CF) layer. These phases will be discussed separately in the following 4.1, 4.2, 4.3 sections.

### 4.1 Body Intensity layer

First, in the BI layer, the activities are separated into two sets called body-fixed set and body-unfixed set depending on the intensity of activities. We mainly focus on the movement of thigh, which can be measured by the accelerometer data recorded by smartphone. Nine kinds of activities are considered as body-fixed activities including eating, typing, drinking, waving hands, reading, stretching, brushing teeth, sleeping, and washing hands. Thigh positions in these nine activities remain static in chair or show slight fluctuation (Figure 2e-h). On the other hand, data from body-unfixed activity set including running, walking, ascending stairs and descending stairs (Figure 2a-d) vibrate much than the other set. Activities in this set show significant difference in the data captured from the thigh; therefore, they could be distinguished from the body-fixed set based on a threshold which is determined through experiments.

This layer utilizes the inherent characteristics of two major categories of activities. The characteristics are reflected by the movement of thigh and they can be captured by the smartphone in the pocket of pants.

Two features, energy and variance, are tested to divide the thirteen activities into body-fixed set and body-unfixed set. Both of them are effective in discriminating low intensity activities from moderate or high intensity activities. Thus, the sum of energy and variance of three axes are extracted. By comparing the number of misclassified samples in experiment based on these two features, the better one will be chosen as the feature to generate threshold which is given in *Evaluation* section.

### 4.2 SVM Classification layer

The SVM Classification (SC) layer is the core layer of the proposed method. In the SC layer, two SVM classifiers are trained

**Table 1** Activities

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Long Duration Activities** | Eating | Typing | Brushing teeth | Reading | Walking | Running | Sleeping |
| **Short Duration Activities** | Drinking | Stretching | Washing hands | Waving hands | Ascending stairs | Descending stairs | |

**Table 2** Features extracted from each sample of raw acceleration data.

| No. | Feature Description |
|---|---|
| **1** | Mean value for each axis(x, y, z) of phone |
| **2** | Mean value for each axis(x, y, z) of watch |
| **3** | Variance for each axis(x, y, z) of phone |
| **4** | Variance for each axis(x, y, z) of watch |
| **5** | Sum of the 3-axis variance of phone |
| **6** | Energy value for each axis(x, y, z) of phone |
| **7** | Energy value for each axis(x, y, z) of watch |
| **8** | Sum of the 3-axis energy value of phone |
| **9** | Collection time label |
| **10** | Collection location label |

respectively for body-fixed activities and body-unfixed activities. Based on the output of BI layer, the body-fixed and body-unfixed SVM classifiers are trained with the twelve-dimensional feature vector, as presented in (4).

$$feature = (mean_x^p, mean_y^p, mean_z^p, var_x^p, var_y^p, var_z^p, mean_x^w, mean_y^w, mean_z^w, var_x^w, var_y^w, var_z^w) \quad (4)$$

where $mean_x^p, mean_x^w$ are the mean value for $x$ axis from phone and watch, $var_x^p$, $var_x^w$ are the variance for $x$ axis from phone and watch. In order to fuse the contextual information, SVM classifier with probability outputs is trained. Probabilities of 13 activities given training data $P(A = Walking|D)$, $P(A = Running|D)$, ..., $P(A = Washinghands|D)$ are generated. For the body-fixed sample, it is possible that a sample belongs to body-unfixed activities is close to zero and vice versa.

## 4.3 Contextual Fusion layer

The Contextual Fusion (CF) layer is designed to refine the output of the SC layer by fusing contextual information, since there is a certain relationship between activities and their environment. A SVM and naïve Bayes fusion algorithm is proposed to use the probability result generated in SC layer. Different from the naïve Bayes algorithm, the fusion algorithm obtains the probability of activity given sensor data $P(A|D)$ from the results of SVM. In this stage, a location list and a time list are defined, which record several presetting locations and time periods. Each sample will be given a location value and a time value within the location list and time list. The *time* values are obtained by dividing one day into several periods and the sample collection time is matched into one of the periods, while the *location* values are obtained by the proposed WiFi-assisted labeling method. The descriptions of the two lists and the corresponding methods are as follows.

1) *Open location list and the WiFi-assisted GPS labeling method. Basically, five location labels (outdoors, dormitory, canteen, office, unknown indoors)* are preset as the attributes in the initial location list. The algorithm first assigns each collected sample a GPS coordinate either from the GPS satellite when subject is outdoors or from WiFi when indoors. To convert GPS coordinates into location labels, a WiFi-assisted GPS labeling method which marks the collected samples with location labels is proposed.

We first deal with the WiFi-assisted indoor locations. In our case, WiFi network of three indoor locations is available, including dormitory, canteen and office. We assumed that the devices can connect to the known indoor network automatically. Because the satellite signal cannot be acquired in these indoor locations, the GPS coordinates can only be obtained from WiFi. However, there may be some offsets in the GPS coordinates caused by the WiFi network and it may changes every time when reconnection occurs; therefore, circular areas are recorded as the corresponding coordinates of these locations instead of precise coordinates. The center coordinates of these circles are updated constantly and calculated as the mean value of all the previous center coordinates. Radius of the circle is 22 meters which is slightly larger than the offsets obtained from measuring the samples. Any GPS coordinates within the circular area are considered as *dormitory, office* or *canteen* (Figure 3).

When the device accesses an unknown WiFi and gets a coordinate out of any circular areas, the label *unknown indoors* is given to the current sample. The sample also gets the *unknown indoors* where no coordinates exists, indicating an unfamiliar location without satellite signal or WiFi network. The labeling method is illustrated in Figure 4.

It should be noted that the location list is open. Any place that provides meaningful and relatively fixed location information can be added to the list. More WiFi available indoor locations such as coffee house or gymnasium can be considered. For the outdoor cases, outdoor athletic fields and basketball court can be added to enrich the contextual information.

2) *Time list and the time labeling method.* Similar to the location list, six time periods (*morning, forenoon, noon, afternoon, evening, night*) are defined in the time list (Table 3). The collection time is recorded when collecting data, then the algorithm

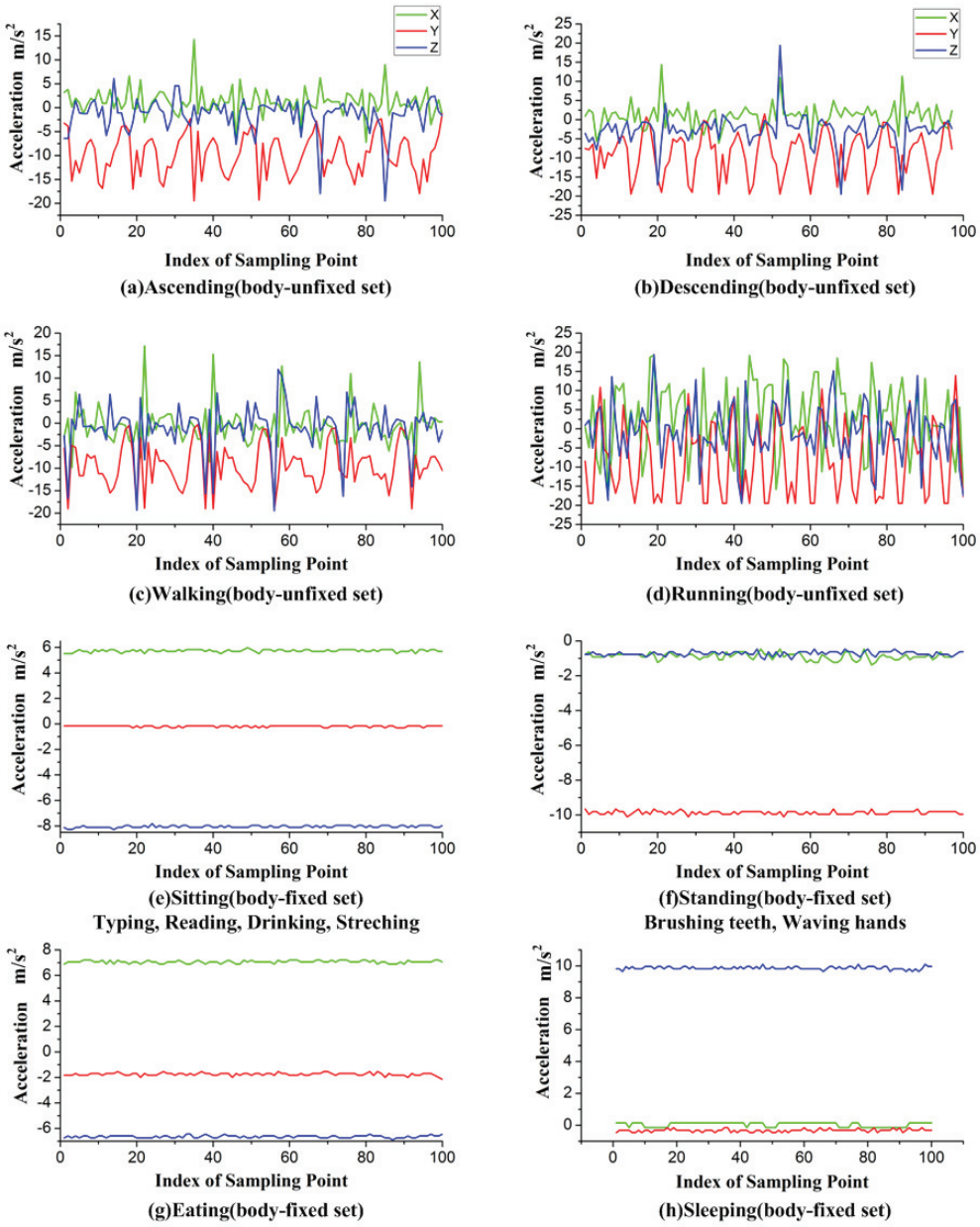


**Figure 2** Accelerometer data of the smartphone from eight activities (a-h).

will match the colleting time with the time list and assign the sample corresponding time label. For example, if a sample is collected at 15:23:20, then it would be labeled as *afternoon*.

Then the *location* and *time* probability matrices are calculated by the training samples, which are described below.

3) *Probability Table of Location (PToL)*. A probability table formed by $13 \times 4$ probability values is generated from the training samples. The element of this table noted as $P(L|A)$ represents the probability of different locations given a certain activity. Values of elements from each row add up to 1. It can be seen as a prior experience obtained by analyzing the location label of training samples.

4) *Probability Table of Time (PToT)*. In common with the probability table of location, a probability table formed by $13 \times 6$ probability values is also obtained by the training samples. The element $P(T|A)$ of this table represents the probability of a time period given a certain activity.

A naïve Bayes model illustrated in Figure 5 is used for fusion. Three kinds of data including GPS coordinates, time, and accelerometer data are fused in the model. The goal of the algorithm is to obtain the recognition activity label $A$ that has the largest probability given training data $D$, time label $T$, location label $L$. The equation is derived as follows.

$$\begin{aligned} &Activity \\ &= argmax_{a\in A}(P(A|D) \times P(A|T) \times P(A|L)) \\ &= argmax_{a\in A}\left(P(A|D) \times \frac{P(T|A) \times P(A)}{P(T)} \right. \\ &\qquad \left. \times \frac{P(L|A) \times P(A)}{P(L)}\right) \\ &= argmax_{a\in A}(P(A|D) \times P(T|A) \times P(A) \\ &\qquad \times P(L|A) \times P(A)) \end{aligned} \tag{5}$$

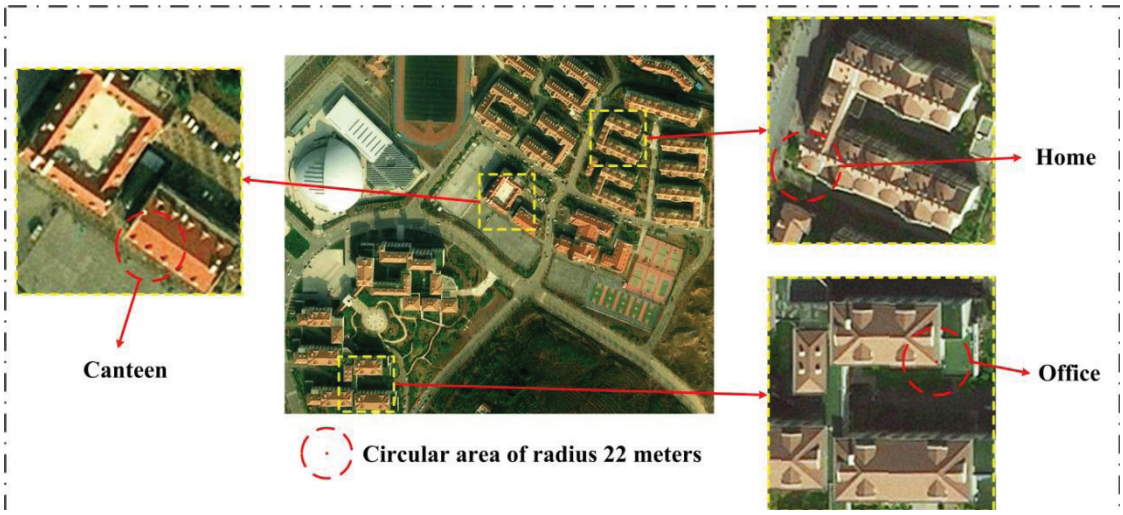


**Figure 3** Circular areas recorded in location list.

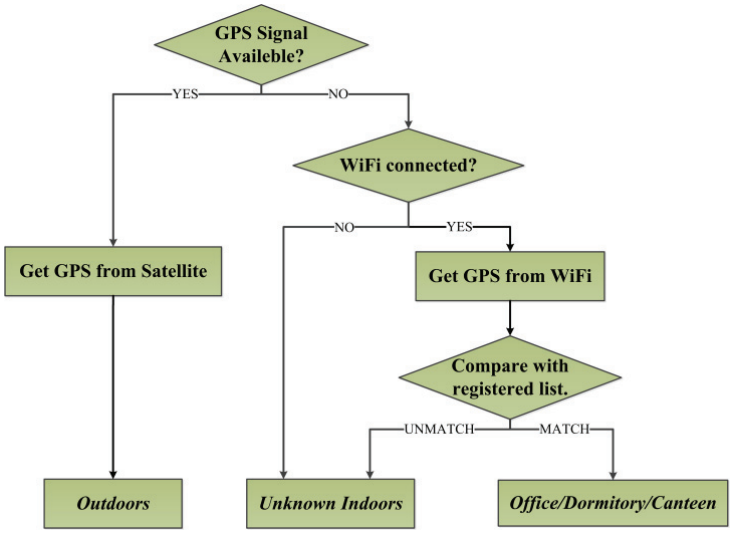


**Figure 4** WiFi-assisted GPS labeling method.

**Table 3** Time list

| | Morning | Forenoon | Noon | Afternoon | Evening | Night |
|---|---|---|---|---|---|---|
| **Time Period** | 6:30–8:00 | 8:00–11:00 | 11:00–13:30 | 13:30–19:00 | 19:00–22:30 | 22:30–6:30 |

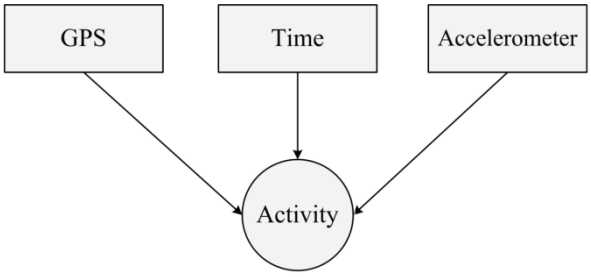


**Figure 5** The naïve bayes model.

where $P(A)$ is a fixed value that can be obtained through the training set. $P(A|D)$ is the probability of activity given training data, which is calculated by SVM classifier in SC layer. $P(T|A)$ and $P(L|A)$ are the probability of time and location that is given in CF layer.

# 5. EVALUATIONS

## 5.1 Data collection

As mentioned in Section 1, we utilized a smartwatch and a smartphone to collect data. The watch is Omate True Smart with various sensors including a GPS sensor, a light sensor, a tri-axial accelerometers, a gyroscopes, etc. The phone is Samsung Galaxy mini, which is also embedded with these sensors. Both of them are powered by Android system, the range of the tri-axial accelerometers is $\pm 2$g and the sampling rate of it is set to 17Hz.

Ten subjects (Six males, four females) were involved in our experiments. The ages of subjects are from 23 to 31. They carried the smartphone in their right leg pants pocket and the smartwatch on their left wrist, as shown in Figure 6. Both of the two devices had installed the data collection application. The two devices communicate with each other via Bluetooth and the smartwatch is responsible for receiving the start and end commands from subjects. Labels of collected data were annotated manually. Data of each long duration activity process between *start* and *end* command was transformed into 7-second samples. For the short duration activities, the app on smartphone and smartwatch only receive the *start* command and record the first 7 seconds data each time. The 7-second data is used as sample directly without further dividing.

## 5.2 Determination of thresholds for dividing activity sets

In order to evaluate the effectiveness of the variance and the energy feature in distinguishing body-fixed set and body-unfixed set, both thresholds of the two features need to be obtained by measuring the misclassified numbers among 7536 samples. One third of these samples (*generating samples*) are used to generate the threshold while the rest (*assessment samples*) are used to assess it. The sum of variance $sum_v$ of three axes is calculated for each generating sample, along with the sum of energy $sum_e$. The $thres + v$ and $thres_e$ are determined when the total number of misclassified samples is minimum. With the growth of $sum_v$ or $sum_e$, the number of misclassified samples in body-fixed set increases while the number of body-unfixed set declines, as shown in Figure 7.

Figure 7(a) shows that, though the curve is close to flat, the number of misclassified reducing to the minimum 478 when the $sum_e$ is 13700. In contrast with the energy feature, there is only 13 (0.26%) misclassified samples at the lowest point among all 5024 assessment samples divided by variance when the $sum_v$ is around 11, as shown in Figure 7(b). In conclusion, the variance provides a simple and effective way to divide the two sets, which is adopted as the dividing method in the BI layer.

## 5.3 Generation of PToT & PToL

We first divide all the samples into ten sets. In each validation nine sets are utilized as training samples and the rest one as testing samples. *Probability table of time* (PToT) and *probability table of location* (PToL) are generated for each 10-fold cross-validation. Table 4 and Table 5 present one of the validation's probabilities.

In the probability table of location, $P(L|A) = 0$ means the activity $A$ has never happened in this location. Take the eating as an example, subjects eat in the canteen in most cases (a high probability of 0.894). Subjects may also have meals in dormitory, office or unknown indoor places when they bring them there. Subjects do not eat outdoors in these training sets, thus $P(Outdoors/Eating) = 0$. In fact, for the indoors cases, more WiFi networks could be recorded in the location list such as convenience store and coffee shop to enrich the contextual information.

Similar to the PToL, $P(T|A) = 0$ means the activity $A$ has never happened during the time period $T$. Take the sleeping activity as an example, subjects sleep at night in most cases (a high probability of 0.7111). Subjects may also take a rest at noon when they finished the tiring work. They may sleep late on weekend morning. Thus noon and morning are the other two periods that sleeping happens. Subjects do not sleep in the rest periods in these training sets, thus other probabilities are recorded as zero in the table.

## 5.4 Performance evaluation of the context-aware hierarchical method

Performance of the context-aware hierarchical method was estimated using 10-fold cross-validation. The result is shown in Table 6, along with the confusion matrix in Figure 8. The proposed method achieves good results in both body-fixed activity set and body-unfixed activity set. Red frames in the confusion matrix represent the intersection of the two activity sets that divided in BI layer. It is verified that most samples in one of the two sets will not confused with samples in the other set. Then for the body-fixed activities, the smartwatch plays an important role because the accelerometer data from watch provide enough details for classifying activities involving hands. For the body-unfixed part, these activities show obvious characters in variance and mean of accelerometers. Moreover, for some contextual activity such as eating and sleeping, the contextual information improves the performance. However, the result of washing hands is not very good. It is possible that this activity has a similar mean and variance with other activities and it is confused with brushing since they may both occur at the dormitory in the morning.

## 5.5 Comparison between hierarchical method and single level method

To compare the hierarchical method with the single level one, totally three SVM classifiers are trained for body-fixed set, body-unfixed set, and the whole activity set respectively. All SVM classifiers use RBF kernel function. The performances of three


**Figure 6** Data collection devices.

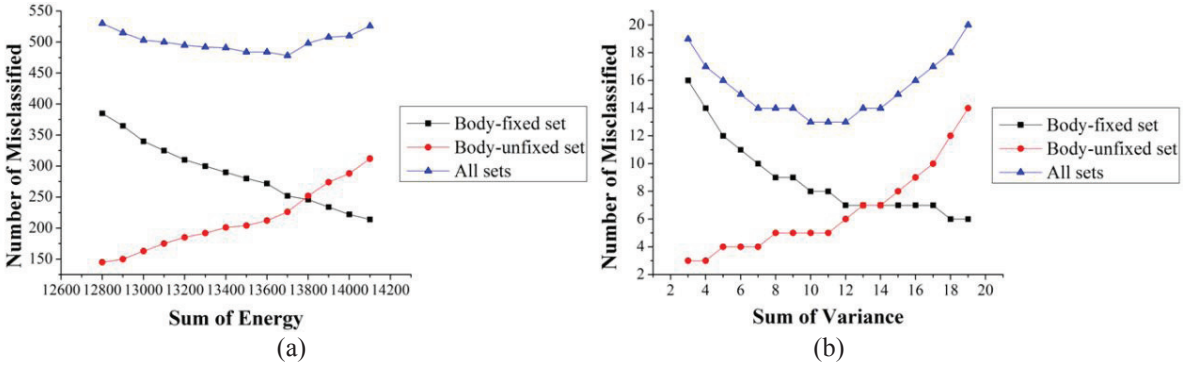


**Figure 7** Number of misclassified samples.

**Table 4** Probability Table of Location (PToL).

| | Unknown Indoors | Outdoors | Dormitory | Office | Canteen |
|---|---|---|---|---|---|
| Running | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 |
| Ascending stairs | 0.980 | 0.020 | 0.000 | 0.000 | 0.000 |
| Descending stairs | 0.970 | 0.030 | 0.000 | 0.000 | 0.000 |
| Walking | 0.092 | 0.849 | 0.019 | 0.008 | 0.032 |
| Eating | 0.007 | 0.000 | 0.081 | 0.018 | 0.894 |
| Typing | 0.011 | 0.000 | 0.108 | 0.882 | 0.000 |
| Drinking | 0.002 | 0.054 | 0.181 | 0.366 | 0.398 |
| Waving hands | 0.040 | 0.590 | 0.145 | 0.132 | 0.093 |
| Reading | 0.000 | 0.000 | 0.327 | 0.673 | 0.000 |
| Stretching | 0.000 | 0.000 | 0.655 | 0.345 | 0.000 |
| Brushing teeth | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 |
| Sleeping | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 |
| Washing hands | 0.062 | 0.000 | 0.875 | 0.000 | 0.063 |

classifiers were estimated using 10-fold cross-validation. The F1-measure of single level classifier is 0.904 on average. The result of classifier for body-fixed set is 0.942, and the body-unfixed set is 0.918, which are both higher than 0.904. We then use the proposed hierarchical method, the F1-measure of which is 0.937 for the overall samples including all activities, improving the average F1-measure by 0.033. This can be explained that the BI layer divided the body-fixed set and body-unfixed set effectively (only 0.26% of all samples) and there are few misclassified samples in the crossed area. The results are illustrated in Figure 9. The F1-measure of most activities are close, and some activities, like running (0.251↑) and brushing (0.090↑), are improved significantly. Although some activities F1-measure declined, it is not obvious (Drinking 0.005↓).

**Table 5** Probability Table of Time (PToT).

| | Morning | Forenoon | Noon | Afternoon | Evening | Night |
|---|---|---|---|---|---|---|
| **Running** | 0.2517 | 0.0000 | 0.0000 | 0.4204 | 0.3279 | 0.0000 |
| **Ascending stairs** | 0.0000 | 0.3713 | 0.3431 | 0.2856 | 0.0000 | 0.0000 |
| **Descending stairs** | 0.3213 | 0.0000 | 0.3212 | 0.3575 | 0.0000 | 0.0000 |
| **Walking** | 0.1229 | 0.1890 | 0.1960 | 0.3100 | 0.1821 | 0.0000 |
| **Eating** | 0.2815 | 0.0290 | 0.3314 | 0.0372 | 0.3209 | 0.0000 |
| **Typing** | 0.0000 | 0.4542 | 0.0000 | 0.4598 | 0.0860 | 0.0000 |
| **Drinking** | 0.1015 | 0.2480 | 0.2217 | 0.2293 | 0.1995 | 0.0000 |
| **Waving hands** | 0.2792 | 0.1672 | 0.2824 | 0.2475 | 0.0237 | 0.0000 |
| **Reading** | 0.0000 | 0.3972 | 0.0000 | 0.4502 | 0.1526 | 0.0000 |
| **Stretching** | 0.4051 | 0.1316 | 0.3112 | 0.0920 | 0.0601 | 0.0000 |
| **Brushing teeth** | 0.4612 | 0.0000 | 0.0562 | 0.0000 | 0.4692 | 0.0134 |
| **Sleeping** | 0.0786 | 0.0000 | 0.2103 | 0.0000 | 0.0000 | 0.7111 |
| **Washing hands** | 0.2857 | 0.0539 | 0.3008 | 0.0542 | 0.2810 | 0.0271 |

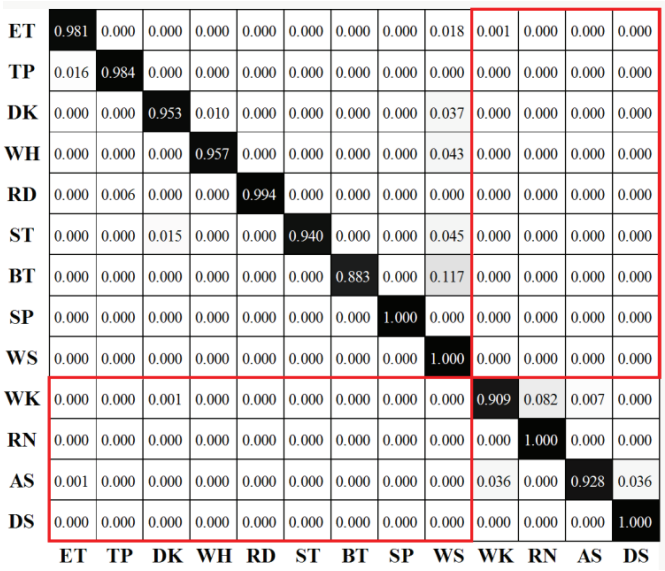


**Figure 8** Confusion matrix of proposed method.

**Table 6** Result of proposed methods.

| Activity | Abbreviation | Recall | Precision | F1 |
|---|---|---|---|---|
| Eating | ET | 98.11% | 99.24% | 0.987 |
| Typing | TP | 98.43% | 100.00% | 0.992 |
| Drinking | DK | 95.32% | 98.50% | 0.969 |
| Waving hands | WH | 95.71% | 100.00% | 0.978 |
| Reading | RD | 99.44% | 100.00% | 0.997 |
| Stretching | ST | 93.98% | 100.00% | 0.969 |
| Brushing teeth | BT | 88.28% | 100.00% | 0.938 |
| Sleeping | SP | 100.00% | 100.00% | 1.000 |
| Washing hands | WS | 100.00% | 73.13% | 0.845 |
| Walking | WK | 90.94% | 99.21% | 0.949 |
| Running | RN | 100.00% | 91.02% | 0.953 |
| Ascending stairs | AS | 92.86% | 96.30% | 0.945 |
| Descending stairs | DS | 100.00% | 97.14% | 0.985 |

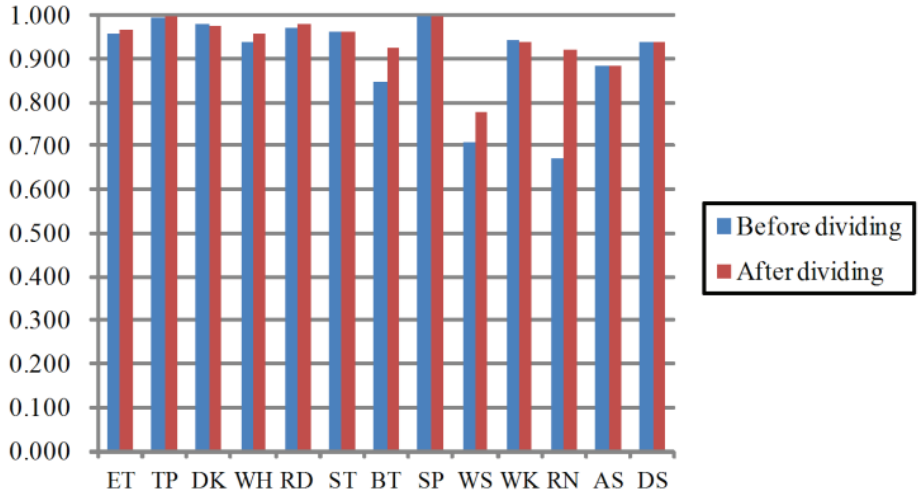


**Figure 9** F1-measurement results before and after dividing.

### 5.6 Comparison between Smartphone Only and Smartphone with Smart Watch

The smartwatch plays an important role in recognizing all 13 activities. To illustrate it more intuitively, SVM classifiers with and without data from the watch are tested. Both of the two classifiers uses hierarchical scheme during their training process. The performances of them were estimated using 10-fold cross-validation. The results are shown in Figure 10. The overall F1-measure of method without smartwatch is 0.763, while the one with watch is 0.937, which is improved by 0.174. The data from smartwatch has small effect on some body-unfixed activities such as walking, ascending and descending, because the phone in pants pocket provides enough information in recognizing these activities. But for the body-fixed activities, like drinking, stretching, waving hands, the performance are improved significantly with the help of the smartwatch. Since the smartwatch capture additional movement data from hand, many hand-involved activities could be recognized.

### 5.7 Performance evaluation of contextual information fusion

In the final stage, the contextual information in CF layer is fused for classifiers. The performance of classifier was also estimated using 10-fold cross-validation. The result is illustrated in Figure 11. On the basis of the SC layer, the CF layer further increases the average F1-measure by 0.027.

As can be seen from the Figure 11, the CF layer performs well on those activities that have distinct contexts. Washing hands and ascending stairs activities are the top two that benefit most from the CF layer. This is because subjects usually washing hands in dormitory within a certain time period. Ascending stairs occur in the building in most cases, with the location label *unknown indoors*, which helps distinguishing them from running that occur outdoors. Beside these two activities, performances of most of the rest activities are improved in CF layer.

Through the test samples, we found that the contextual information did not help to reclassify the walking samples correctly. The walking samples are confused with outdoor running and indoor ascending due to the lower location probability of indoor walking. Moreover, the walking samples do not take any advantage by the time label, for example, both walking and running may happen at any daytime.

## 6. COMPARISON BETWEEN THE PROPOSED METHOD AND EXISTING METHODS

Finally, we compared the proposed method with other classification methods including Neural Network, C4.5 Decision Tree and Support Vector Machine.

For the Neural Network and Decision Tree, the corresponding hierarchical classifier and the hierarchical classifier fusing time and location are also tested to verify the effectiveness of the hierarchical scheme and the contextual information. It can be seen that the hierarchical scheme fusing contextual information are effective for both Neural Network classifier and C4.5 Decision Tree. Through this method, the F1 value of the hierarchical Neural Network fusing contextual information is 0.954, improving 0.146 comparing with the single-layer Neural Network. Then, the F1 value of the hierarchical C4.5 Decision Tree fusing contextual information is 0.945, improving 0.044 comparing with the single-layer C4.5 Decision Tree. For the Support Vector Machine, three methods are tested, including single-layer SVM classifier, SVM fusing time only, and SVM classifier fusing location only. All the results are illustrated in Table 7. It is demonstrated that the proposed hierarchical scheme fusing contextual information is effective for all the evaluated methods (Neural Network, C4.5 Decision Tree, Support Vector Machine), and the hierarchical SVM fusing contextual information method performs best among them.

## 7. CONCLUSIONS

In this paper, a context-aware hierarchical approach is proposed to recognize 13 elementary activities only with a wearable watch

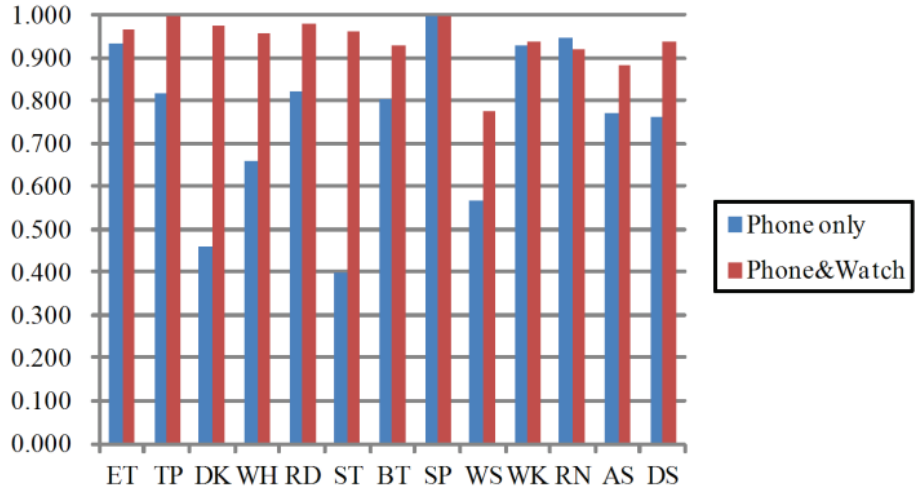


**Figure 10** F1-measurement results with and without smartwatch.

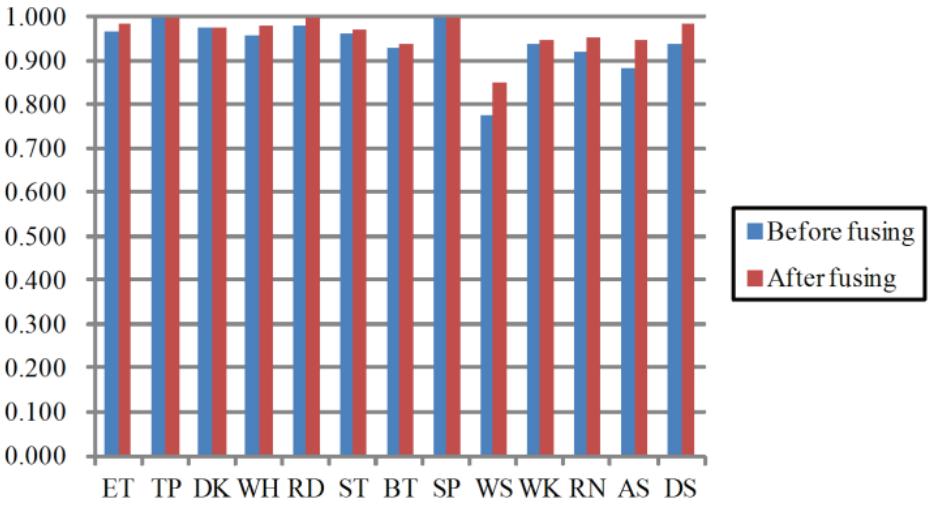


**Figure 11** F1-measurement results before and after fusing contextual information.

**Table 7** Performance comparison of NN, C4.5 DT, SVM and the proposed method.

| Algorithms | Recall | Precision | F1 |
|---|---|---|---|
| Neural Network | 83.05% | 78.74% | 0.808 |
| Hierarchical Neural Network | 91.17% | 92.36% | 0.918 |
| Hierarchical NN Fusing Time and Location | 95.46% | 95.37% | 0.954 |
| C4.5 Decision Tree | 89.91% | 90.25% | 0.901 |
| Hierarchical C4.5 Decision Tree | 89.85% | 89.56% | 0.897 |
| Hierarchical DT Fusing Time and Location | 94.71% | 94.26% | 0.945 |
| Support Vector Machine | 88.98% | 91.82% | 0.904 |
| Hierarchical SVM Fusing Time | 91.12% | 94.41% | 0.927 |
| Hierarchical SVM Fusing Location | 94.68% | 95.92% | 0.953 |
| Proposed Method | 96.42% | 96.47% | 0.964 |

and a smartphone. The proposed approach first divides activities into two parts according to the intensity of activities, and each part is classified respectively. Furthermore, a WiFi-assisted GPS labeling method and a time labeling method are proposed to utilize contextual information, and a naÃ¯ve Bayes model is presented for the fusion. The WiFi-assisted GPS labeling method utilized WiFi positioning to deal with the known indoors location such as dormitory, canteen and office. A lot of comparative experiments are designed and conducted to evaluate the proposed method. First, dividing activity sets is a good way to deal with a variety of activities, because it improves the accuracy of each activity. And the experiment results show that the sum of three-axis variance provides a simple but effective method. Second, the smartwatch is helpful in recognizing some body-fixed activi-

ties because the introduction of the hand movement data. Third, contextual information including location and time is useful to refine the output of SVM.

Our work provides a new method for fusing contextual information into activity recognition. We are considering several directions for future work. One of them is extending the contextual information by exploring more sensors in mobile devices. For example, the microphone could be used to perceive occasions of activities by analyzing the current noise, and the velocity provided by the GPS could be used to determine whether the subject is in a car. With the integration of more multimodal context data, a complete contextual fusion framework can be built for Android devices, thus more complex activities or human behaviors, such as playing basketball or giving a lecture, will be recognized.

## Acknowledgements

This work is supported in part by National Natural Science Foundation of China (No.61602430, No.61672475, No.61402428); Aoshan Science and Technology Innovation Program of QNLM (2016ASKJ07). The authors would thank anonymous reviewers for their valuable comments.

## REFERENCES

1. Gayathri, K. S.; Elias, S.; Ravindran, B. Hierarchical activity recognition for dementia care using markov logic network. *Pers. Ubiquitous Comput*. **2015**, 19, 271–285.
2. Avci, A.; Bosch, S. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: *A survey. Architecture of computing systems (ARCS)*, **2010** 23rd international conference on. VDE 2010, 1–10.
3. Hong, Y.-J.; Kim, I.-J.; Ahn, S. C.; Kim, H.-G. Mobile health monitoring system based on activity recognition using accelerometer. *Simul. Model. Pract. Theory* **2010**, 18, 446–455.
4. Wan, J.; O'grady, M. J.; O'hare, G. M. Dynamic sensor event segmentation for real-time activity recognition in a smart home context. *Pers. Ubiquitous Comput.* **2015**, 19, 287–301.
5. Belley, C.; Gaboury, S.; Bouchard, B.; Bouzouane, A. An efficient and inexpensive method for activity recognition within a smart home based on load signatures of appliances. *Pervasive Mob. Comput.* **2014**, 12, 58–78.
6. McAvoy, L. M.; Chen, L.; Donnelly, M. P.; Nugent, C. D.; McCullagh, P. J. Ontological characterization and representation of context within smart environments. *Comput. Syst. Sci. Eng.* **2015**, 30, 19–32.
7. Rafferty, J.; Chen, L.; Nugent, C.; Liu, J. Goal lifecycles and ontological models for intention based assistive living within smart environments. *Int. J. Comput. Syst. Sci. Eng.* **2015**, 30, 1–14.
8. Zhao, C.; He, J.; Zhang, X.; Qi, X.; Chen, A. Recognition of driving postures by nonsubsampled contourlet transform and k-nearest neighbor classifier. *Comput. Syst. Sci. Eng.* **2015**, 30, 233–241.
9. Bulling, A.; Blanke, U.; Schiele, B. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.* **2014**, 46, 33.
10. Lara, O. D.; Labrador, M. A. A survey on human activity recognition using wearable sensors. *Commun. Surv. Tutorials, IEEE* **2013**, 15, 1192–1209.
11. Bao, L.; Intille, S. S. Activity recognition from user-annotated acceleration data. In *Pervasive computing*; Springer, **2004**; pp. 1–17.
12. Liao, L.; Fox, D.; Kautz, H. Hierarchical conditional random fields for GPS-based activity recognition. In *Robotics Research*; Springer, **2007**; pp. 487–506.
13. Riboni, D.; Bettini, C. Cosar: hybrid reasoning for context-aware activity recognition. *Pers. Ubiquitous Comput.* **2011**, 15, 271–289.
14. Khan, A. M.; Lee, Y.-K.; Lee, S. Y.; Kim, T.-S. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *Inf. Technol. Biomed. IEEE Trans*. **2010**, 14, 1166–1172.
15. Long, X.; Yin, B.; Aarts, R. M. Single-accelerometer-based daily physical activity classification. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*; **2009**; pp. 6107–6110.
16. Mannini, A.; Intille, S. S.; Rosenberger, M.; Sabatini, A. M.; Haskell, W. Activity recognition using a single accelerometer placed at the wrist or ankle. *Med. Sci. Sports Exerc*. **2013**, 45, 2193–2203.
17. Lee, M.-W.; Khan, A. M.; Kim, T.-S. A single tri-axial accelerometer-based real-time personal life log system capable of human activity recognition and exercise information generation. *Pers. Ubiquitous Comput*. **2011**, 15, 887–898.
18. Cleland, I.; Kikhia, B.; Nugent, C.; Boytsov, A.; Hallberg, J.; Synnes, K.; McClean, S.; Finlay, D. Optimal placement of accelerometers for the detection of everyday activities. *Sensors* **2013**, 13, 9183–9200.
19. Gao, L.; Bourke, A. K.; Nelson, J. Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems. *Med. Eng. Phys.* **2014**, 36, 779–785.
20. Han, M.; Bang, J. H.; Nugent, C.; McClean, S.; Lee, S. A Lightweight Hierarchical Activity Recognition Framework Using Smartphone Sensors. Sensors 2014, 14, 16181–16195.
21. Guiry, J. J.; van de Ven, P.; Nelson, J.; Warmerdam, L.; Riper, H. Activity recognition with smartphone support. *Med. Eng. Phys.* **2014**, 36, 670–675.
22. Tapia, E. M.; Intille, S. S.; Haskell, W.; Larson, K.; Wright, J.; King, A.; Friedman, R. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In *Wearable Computers, 2007 11th IEEE International Symposium on*; **2007**; pp. 37–40.
23. Maurer, U.; Rowe, A.; Smailagic, A.; Siewiorek, D. Location and activity recognition using eWatch: A wearable sensor platform. In *Ambient Intelligence in Everyday Life*; Springer, **2006**; pp. 86–102.
24. Ermes, M.; Parkka, J.; Cluitmans, L. Advancing from offline to online activity recognition with wearable sensors. In Engineering in Medicine and Biology Society, 2008. EMBS 2008. *30th Annual International Conference of the IEEE*; **2008**; pp. 4451–4454.
25. Zeng, M.; Nguyen, L. T.; Yu, B.; Mengshoel, O. J.; Zhu, J.; Wu, P.; Zhang, J. Convolutional neural networks for human activity recognition using mobile sensors. In Mobile Computing, Applications and Services (MobiCASE), *2014 6th International Conference on*; **2014**; pp. 197–205.
26. Yang, J. B.; Nguyen, M. N.; San, P. P.; Li, X. L.; Krishnaswamy, S. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), Buenos Aires, Argentina*; **2015**; pp. 25–31.
27. Ordonez, F. J.; Roggen, D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* **2016**, 16, 115.
28. Chavarriaga, R.; Sagha, H.; Calatroni, A.; Digumarti, S. T.; Tröster, G.; MillÃ¡n, J. del R.; Roggen, D. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognit. Lett.* **2013**, 34, 2033–2042.

29. Frank, K.; Rockl, M.; Nadales, M. J. V.; Robertson, P.; Pfeifer, T. Comparison of exact static and dynamic bayesian context inference methods for activity recognition. In Pervasive Computing and Communications Workshops (PERCOM Workshops), *2010 8th IEEE International Conference on*; **2010**; pp. 189–195.
30. He, Z.-Y.; Jin, L.-W. Activity recognition from acceleration data using AR model representation and SVM. In *Machine Learning and Cybernetics, 2008 International Conference on*; **2008**; Vol. 4, pp. 2245–2250.
31. Maurer, U.; Smailagic, A.; Siewiorek, D. P.; Deisher, M. Activity recognition and monitoring using multiple sensors on different body positions. In Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. *International Workshop on*; **2006**; p. 4–pp.
32. Li, Z.; Wei, Z.; Huang, L.; Zhang, S.; Nie, J. Hierarchical Activity Recognition Using Smart Watches and RGB-Depth Cameras. *Sensors* **2016**, 16, 1713.
33. Martín, H.; Bernardos, A. M.; Iglesias, J.; Casar, J. R. Activity logging using lightweight classification techniques in mobile devices. *Pers. ubiquitous Comput*. **2013**, 17, 675–695.
34. Kwapisz, J. R.; Weiss, G. M.; Moore, S. A. Activity recognition using cell phone accelerometers. *ACM SigKDD Explor. Newsl.* **2011**, 12, 74–82.
35. Khan, A. M.; Lee, Y.-K.; Lee, S. Y.; Kim, T.-S. Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis. In Future Information Technology (FutureTech), *2010 5th International Conference on*; **2010**; pp. 1–6.
36. Ronao, C. A.; Cho, S.-B. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Syst. Appl.* **2016**, 59, 235–244.
37. Lara, O. D.; Pérez, A. J.; Labrador, M. A.; Posada, J. D. Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive Mob. Comput.* **2012**, 8, 717–729.