Smartphone Sensor Based Human Activity Recognition Using Feature Fusion and Maximum Full A Posteriori

Zhenghua Chen, Chaoyang Jiang, Shili Xiang, Jie Ding, Min Wu, and Xiaoli Li

***Abstract*—Human activity recognition (HAR) using smartphone sensors has attracted great attention, due to its wide range of applications. A standard solution for HAR is to firstly generate some features defined based on domain knowledge (*handcrafted* features), and then to train an activity classification model based on these features. Very recently, deep learning with *automatic* feature learning from raw sensory data has also achieved great performance for HAR task. We believe that both the *handcrafted* features and the *learned* features may convey some unique information which can complement each other for HAR. In this paper, we firstly propose a feature fusion framework to combine handcrafted features with automatically learned features by a deep algorithm for HAR. Then, taking the regular dynamics of human behaviour into consideration, we develop a *maximum full a posterior* (MFAP) algorithm to further enhance the performance of HAR. Our extensive experimental results show the proposed approach can achieve superior performance comparing with state-of-the-art methodologies across both a public dataset and a self-collected dataset.**

***Index Terms*—HAR, smartphone sensors, deep learning, feature fusion, MFAP**

I. Introduction

Human activity recognition (HAR) is of great importance for many applications in heath-care services, smart homes and pervasive and mobile computing [1], [2]. With the development of computer vision techniques, camera-based HAR has been well developed [3]. However, it can only monitor a specific space with adequate illumination condition. In addition, it suffers from privacy concerns. Wearable sensors, such as accelerator and gyroscope, are also popular for HAR [4], [5]. However, they require special hardware to be worn by users, which is obviously inconvenient. In the past decade, smartphones become more and more powerful with many sensors embedded, including accelerator, gyroscope, barometer, temperature sensor, etc. Since most of people carry smartphones in their daily life, smartphones based HAR will thus be a practical option [6], [7].

This work is supported by MND (Ministry of National Development) Singapore, Sustainable Urban Living Program, under the grant no. SUL2013-5, and the Beijing Institute of Technology Research Fund Program for Young Scholars. *(Corresponding Author: Chaoyang Jiang; Min Wu.)*

Zhenghua Chen, Shili Xiang, Jie Ding, Min Wu, and Xiaoli Li are with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, Sinagpore 138632 (e-mail: chen0832@e.ntu.edu.sg, sxiang@i2r.a-star.edu.sg, ding_jie@i2r.a-star.edu.sg, wumin@i2r.a-star.edu.sg, xlli@i2r.a-star.edu.sg).

Chaoyang Jiang is with the Science and Technology on Vehicle Transmission Laboratory School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China 100081 (e-mail: cjiang@bit.edu.cn).

Recently, smartphone sensor based HAR has been developed, which can be generally divided into two categories, namely, shallow and deep algorithms. Specifically, shallow algorithms consist of two steps: feature extraction and activity inference [2], [8]. Since the *raw* smartphone sensor data is not well representative for distinct activities, a standard procedure is thus to extract some informative features, also known as feature extraction/engineering. For instance, the magnitude of acceleration should be helpful in separating different activities such as walking and running. As such, some defined statistical features known as handcrafted features will be firstly extracted from the raw smartphone sensor data. Note that, these handcrafted features are also automatically generated by programs which are written based on their definitions. Some machine learning algorithms, such as neural networks, support vector machines and random forest can be then applied with the handcrafted features to identify different human activities. Deep algorithm based HAR, on the other hand, is one step approach which can automatically learn representative features from the raw sensory data for HAR without human intervention, as well as perform activity inference simultaneously [9], [10], [11].

We observe that both shallow learning algorithms with handcrafted features and deep learning algorithms with automatically learned features have achieved great successes for the task of HAR [12], [13]. We believe that both handcrafted features and automatically learned features by deep algorithms may convey unique information which can complement each other to boost the performance of smartphone sensor based HAR. In this work, at the first stage, we propose a feature fusion framework to integrate handcrafted features with a deep algorithm, i.e., deep long short-term memory (LSTM), to boost the performance of HAR. At the second stage, considering the dynamics (frequent activity changes) of human behaviour, we propose a *maximum full a posterior* (MFAP) algorithm which exploits all the past information and the current *a posterior* probability obtained from the feature fusion framework to give an optimal estimation of human activities.

The main contributions of this paper are summarized as follows:

- We propose a novel feature fusion framework which can effectively combine handcrafted features with a deep learning algorithm to boost the performance of smartphone sensor based HAR.
- Taking the dynamics of human behaviours into consideration, we formulate a MFAP algorithm which exploits all the past information and the current *a posterior*

information obtained from the feature fusion framework to give an optimal estimation of human activities.
- We use a public dataset and a self-collected dataset to evaluate the effectiveness of the proposed approach. Our comprehensive experimental results demonstrate the proposed approach significantly outperforms existing advanced learning algorithms and the state-of-the-arts.

The remaining of the paper is organized as follows: Section II reviews some related works with handcrafted features and automatically feature learning by deep algorithms for HAR. Section III briefly introduces the handcrafted features and a deep algorithm for automatic feature learning, followed by the proposed feature fusion framework. Section IV presents the proposed MFAP algorithm. Section V first demonstrates the data for evaluation, followed by the experimental setup. Then, the experimental results are presented and discussed. Section VI concludes this work and presents some potential future works.

II. Related works

In this section, some related works for HAR using different learning algorithms are reviewed. We divide this section into two parts: shallow and deep algorithms.

A. Shallow algorithms

For shallow algorithms, they normally consist of feature engineering and activity inference. Since raw smartphone sensor data is noisy and not representative for different human activities, some more informative features can be extracted with domain knowledge. Then, shallow learning algorithms can be performed for HAR with these handcrafted features. For example, Wang et al. investigated the effectiveness of smartphone accelerator and gyroscope for HAR [14]. Firstly, they extracted a large number of statistical features from both time and frequency domains of three-dimensional acceleration and gyroscope. Then, they proposed a hybrid of filter and wrapper method known as FW to select best features from all handcrafted features. Finally, machine learning algorithms, namely, k-nearest-neighbors (KNN) and naive Bayes (NB) were employed to classify different activities. Eastwood and Jayne evaluated different extensions of hyperbox neural network (HNN) which is built upon different modes of learning for HAR [15]. In addition, Anguita et al. proposed a hardware friendly support vector machine (HF-SVM) algorithm based on fix-point arithmetic for HAR using smartphone sensors [12]. The experimental results showed that HF-SVM has a comparable performance to the conventional SVM, but with much less computational complexity. Ronao and Cho presented two-stage continuous hidden Markov models (CHMM) for HAR [16]. The first-stage CHMM was utilized to separate static and dynamic activities. The second-stage CHMM was then applied to identify the exact activity from the two types of activities. In [17], the authors enhanced the sparse random classifier with singular value decomposition (SRC-SVD) for HAR. The SVD was leveraged to construct the random projection matrix for SRC. Seera et al. proposed a hybrid of fuzzy min-max (FMM) neural network and the classification and regression tree (CART) to recognize human activities [18]. In their proposed system, the FMM was mainly used for data incremental learning and the CART was utilized to provide interpretations for the classification.

B. Deep algorithms

Owing to the powerful feature learning ability of deep algorithms, they have achieved remarkable performance for HAR using smartphone sensors. Li et al. presented a sparse auto-encoder (SAE) to automatically learn representative features from raw smartphone accelerator and gyroscope data for the task of HAR [19]. The three-dimensional acceleration, gyroscope and the magnitudes of them are treated as different channels on which the SAE is implemented for feature learning. Ronao and cho presented a convolutional neural network (convnet) which is able to learn representative features from raw smartphone sensor data for HAR [20]. They also explored the use of temporal fast Fourier transform (tFFT) on the raw sensory data with convnet for HAR. In their another work, they attempted to apply handcrafted features as the inputs of convnet instead of the raw smartphone sensor data for HAR [21]. Tao et al. presented an ensemble bidirectional long short-term memory (BLSTM) approach for HAR [22]. They applied the raw sensory data, the magnitude of the raw sensory data and two-directional features as inputs for different BLSTM. Experiments indicate the effectiveness of their proposed approach. In [13], the authors proposed a knowledge distilling strategy which attempts to use well-designed handcrafted features to guide deep algorithms for generalization for smartphone sensor based HAR. A comprehensive survey on deep learning based HAR can be found in [23].

In real applications, both handcrafted features with domain knowledge and automatically learned features by deep algorithms may convey unique information for HAR. In this work, we attempt to build a feature fusion framework to combine these two types of comprehensive features to make good use of all the useful information, which should boost the performance of HAR. Taking the dynamics of human behaviour into consideration, we further improve the performance of HAR by formulating a MFAP algorithm which exploits all the past information with the current *a posterior* information obtained from the feature fusion framework to give an optimal estimation of human activities.

III. The Proposed Feature Fusion framework

In this section, we will first briefly introduce handcrafted features and automatic feature learning, and subsequently elaborate two key innovations in our proposed methods.

A. Handcrafted features

Feature engineering is a widely used technique for data preprocessing, leading to the success of shallow machine learning algorithms [24]. For HAR using smartphone sensors, the raw sensory data is not representative for different human activities. To achieve better performance for HAR, some more

representative features can be extracted based on domain knowledge. For example, the activities of walking and running will yield different magnitudes of acceleration. Thus, the feature of the magnitude of acceleration can be extracted for the separation of these two activities. In addition, the variance of smartphone sensors can be used to distinguish static activities from dynamic ones. As such, some advanced statistical features from time and frequency domains have been shown to be effective for smartphone sensor based HAR [12], which are presented in Table I. All these handcrafted features will be extracted for both three dimensional acceleration and gyroscope of smartphones.

TABLE I
HANDCRAFTED FEATURES

Domain	Features
Time	Mean
	Standard deviation
	Median absolute
	Maximum
	Minimum
	Signal magnitude area
	Average sum of the squares
	Interquartile range
	Signal Entropy
	Autorregresion coefficients
	Correlation coefficient
Frequency	Largest frequency component
	Weighted average
	Skewness
	Kurtosis
	Energy of a frequency interval
	Angle between two vectors

B. Automatic feature learning

Deep learning has achieved great success in many challenging research areas, such as image recognition [25] and nature language processing [26]. The biggest merit of deep learning is the ability of automatic feature learning from raw sensory data without human intervention. For HAR using smartphone sensors, the raw sensory data is typical time series with temporal dependency [27]. While recurrent neural network (RNN) is naturally suitable for time series data, the conventional RNN suffers from the problem of gradient vanishing and exploding, which degrades its performance on the modeling of long term dependencies in sequential data [28]. To solve this problem, Hochreiter and Schmidhuber proposed a new RNN named long short-term memory (LSTM) which attempts to use some memory cells to preserve information for long term dependencies [29].

A typical structure of LSTM can be found in Fig. 1, where x^t is the input at time step t, h^t is the hidden state, C^{t-1} is the memory cell state, w^f, w^i, w^C and w^o are the weights, b^f, b^i, b^C and b^o are the biases, and $\sigma(\cdot)$ and $tanh$ are the *sigmoid* and *tanh* functions, respectively.

In the LSTM network, the first step is to determine which information should be thrown from the previous memory cell state C^{t-1} by using a forget gate, which can be formulated as

$$f^t = \sigma\left(w^f[h^{t-1}, x^t] + b^f\right), \tag{1}$$

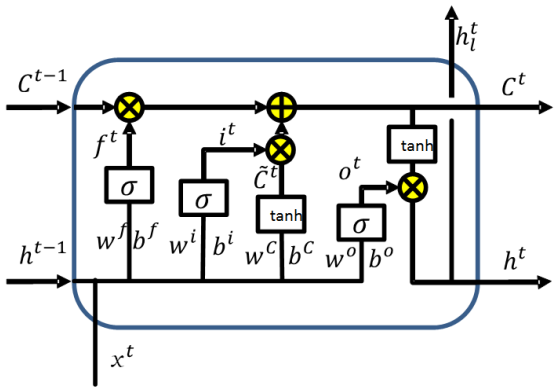


Fig. 1. The structure of the LSTM network

Here, $f^t = 1$ means to keep all the information from the previous step and $f^t = 0$ means to totally remove the information from the previous step. The next step is to determine which new information should be stored based on the current input. It consists of two components. The first component is an input gate to decide what shall be updated. It can be expressed as

$$i^t = \sigma\left(w^i[h^{t-1}, x^t] + b^i\right) \tag{2}$$

The second component produces a candidate state value $\tilde{C}^t$ by using a *tanh* function, shown as

$$\tilde{C}^t = tanh\left(w^C[h^{t-1}, x^t] + b^C\right) \tag{3}$$

After that, the next step is to decide the current state C^t by using the following equation

$$C^t = f^t * C^{t-1} + i^t * \tilde{C}^t \tag{4}$$

Finally, the hidden output h^t is a filtered version of the compressed cell state $tanh\,(C^t)$. The output of the *sigmoid* layer o^t will determine which part of the information will be preserved. It is shown as

$$o^t = \sigma\left(w^o[h^{t-1}, x^t] + b^o\right) \tag{5}$$

The final hidden output $h^t \in \mathbb{R}^d$, where d is the the dimension of the feature, can be expressed as

$$h^t = o^t * tanh\left(C^t\right) \tag{6}$$

Deep architecture has been shown to be effective for representation learning [30]. Therefore, in this work, we *stack multiple LSTM layers*, known as deep LSTM, for deep representation learning in the task of smartphone sensor based HAR. Specifically, the output of i-th LSTM layer will be the input of $(i+1)$-th LSTM layer. As a special case, the input of the *first* LSTM layer is the *raw* sequential smartphone sensor data.

C. Proposed feature fusion

Both the handcrafted features with domain knowledge and the features learned by deep algorithms may contain unique information for HAR. To make good use of these two types of features, we propose a *feature fusion framework* to combine them together for better recognition of human activities using

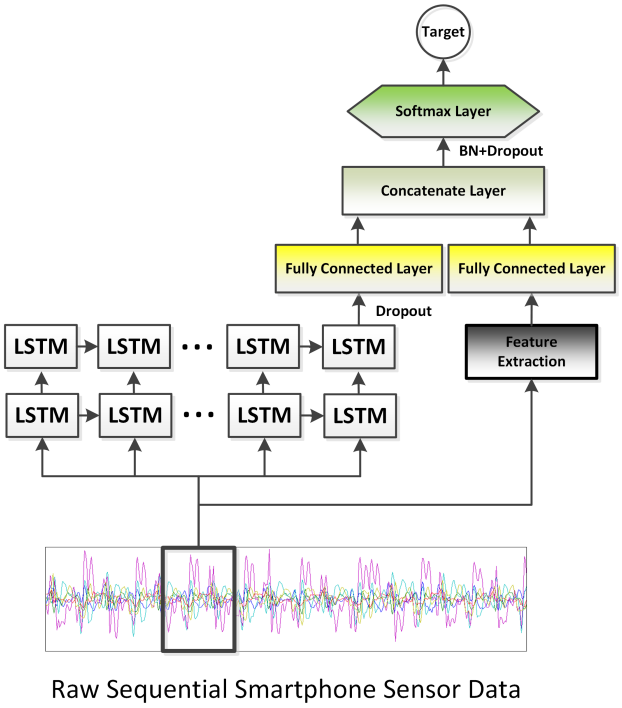


Fig. 2. The proposed feature fusion framework.

smartphone sensors. The proposed feature fusion framework is shown in Fig. 2. Here, we choose the deep LSTM for feature learning, which is naturally suitable for our sequential data analysis problem. The raw sequential smartphone sensor data is fed into two stacked LSTM layers for feature learning. The learned features at the last time instance are fed into a fully connected layer (FCL) to get more abstract features. At the same time, the handcrafted features in Table I, extracted from the raw smartphone sensor data, are fed into another FCL to obtain more abstract features. After that, we combine the two types of features using a concatenate layer. Finally, the combined features are fed into a softmax layer for activity classification.

More specifically, given the smartphone sensor input $\mathbf{o}_t$ which is a window of sensory data, the automatically learned features and the handcrafted features can be expressed as $\mathbf{v}_t = \Phi(\mathbf{o}_t)$ and $\mathbf{h}_t = \Gamma(\mathbf{o}_t)$ respectively, where $\Phi(\cdot)$ is the LSTM based feature learning, and $\Gamma(\cdot)$ is the handcrafted feature extraction based on domain knowledge. Note that, the LSTM is able to encode temporal dependencies within the sample (window) during feature learning. These two types of features can be treated as the processing of the raw sensory data in two distinct perspectives, both of which have been shown to be effective for HAR. The complete feature set is the concatenation of the two types of features, which can be expressed as $\mathbf{l}_t = \mathbf{v}_t \cup \mathbf{h}_t$. This concatenation is able to make full use of these two types of features, which may also lead to a more comprehensive understanding of the raw sensory data. Hence, better performance can be expected. The final outputs of the proposed feature fusion framework are the probabilities of all activities by using the softmax layer on these features, which can be expressed as $softmax(\mathbf{l}_t)$.

The training of the proposed feature fusion framework is to optimize the parameters of the network by using back-propagation algorithm on the training data. Specifically, given training data and targets, the network outputs with the training data are calculated. The errors between the network outputs and the given targets can be obtained, where the gradient of the errors can be used to update network parameters based on gradient-based optimization methods. In this work, we utilize an optimization method of RMSprop which is able to use the magnitude of recent gradients to normalize the gradients [31] for parameter optimization. To prevent over-fitting, some dropout layers and a batch normalization (BN) layer are employed, which are shown in Fig. 2. The dropout rates for the two dropout layers are both set to be 0.5.

After the network has been learned with the training data, the outputs of the proposed feature fusion framework are the probabilities of all activities given the current sensor measurements o_t, which can be expressed as $p(z_t|o_t)$. It is also known as *a posteriori*. Generally, the current activity will be determined based on the maximal probability of *a posteriori*, known as maximum *a posteriori* (MAP) estimation. However, the current human activity should be related to the activity sequence in the past and previous sensor observations, which is not considered by the MAP during estimation. In other words, the LSTM network in the proposed feature fusion framework is only able to encode temporal dependencies within the sample. But it is not able to model the temporal dynamics among samples (activity sequence). To further improve the performance of HAR, we propose a MFAP approach which combines the past information with the current *a posteriori* to give an optimal estimation of human activities.

IV. Maximum Full A Posteriori Estimation

In real life, when performing activities, human normally carries on one activity for a while and then transfer to another activity. This important property should be considered when designing HAR systems. However, to the best of our knowledge, no previous works have exploited this important property of human behaviour. The conventional data-driven approaches attempt to estimate human activities only based on current sensor observations. In this work, to take the dynamics of human behaviour into consideration, we propose a MFAP algorithm which is able to consider the past information and the current *a posterior* information obtained from the proposed feature fusion framework. The MFAP can be formulated as

$$\hat{z}_t = \arg\max_{z_t} \mathbf{p}(z_t|o_{1:t}), \tag{7}$$

where z_t is the human activity at time instance t and $o_{1:t}$ are observations from time instance 1 to t. Here, we make two basic assumptions for HAR using the MFAP algorithm, which are as follows:

1) the state (activity) follows a first-order Markov property, i.e., $p(z_t|z_{t-1}) = p(z_t|z_{1:t-1})$.
2) the current observation of state is conditionally independent from the previous observations, i.e., $p(o_t|o_{1:t-1}, z_t) = p(o_t|z_t)$.

Human normally performs activity sequentially. The current activity usually has high correlation with the activities performed recently and low correlation with the activities performed long ago. This process has been well modeled by a first-order Markov chain [32]. Therefore, we can assume that human activities follow a first-order Markov property, and the first assumption is considered valid. The observation relies on the real human activity. Once the current activity is known, the current observation is independent from the previous observations. Hence, the second assumption which states that the current observation of an activity is conditional independent from the previous observations is also considered valid.

According to Bayes rules, the full *a posterior* probability for HAR, $p(z_t|o_{1:t})$, can be expressed as

$$\begin{aligned} p(z_t|o_{1:t}) &= \frac{p(o_{1:t}|z_t)p(z_t)}{p(o_{1:t})} \\ &= \frac{p(o_t, o_{1:t-1}|z_t)p(z_t)}{p(o_t, o_{1:t-1})} \\ &= \frac{p(o_t|o_{1:t-1}, z_t)p(o_{1:t-1}|z_t)p(z_t)}{p(o_t|o_{1:t-1})p(o_{1:t-1})} \\ &= \frac{p(o_t|o_{1:t-1}, z_t)p(z_t|o_{1:t-1})p(o_{1:t-1})p(z_t)}{p(o_t|o_{1:t-1})p(o_{1:t-1})p(z_t)} \\ &= \frac{p(o_t|z_t)p(z_t|o_{1:t-1})}{p(o_t|o_{1:t-1})} \\ &= \frac{p(z_t|o_t)p(o_t)p(z_t|o_{1:t-1})}{p(z_t)p(o_t|o_{1:t-1})} \end{aligned} \tag{8}$$

Given observations $o_{1:t}$ from time step 1 to t, the probability of $\frac{p(o_t)}{p(o_t|o_{1:t-1})}$ is deterministic, which can be treated as a normalization factor. Therefore, the full *a posterior* probability can be further expressed as

$$p(z_t|o_{1:t}) \propto \frac{p(z_t|o_t)p(z_t|o_{1:t-1})}{p(z_t)} \tag{9}$$

In Equation (9), $p(z_t|o_t)$ is the *a posterior* probability of the human activity. Compared with $p(z_t|o_t)$, full observation information is involved in $p(z_t|o_{1:t})$. Hence, we call the estimation in Equation (7) maximum full *a posterior* (MFAP) estimation. We can find from Equation (9) that the full *a posterior* probability, i.e., $p(z_t|o_{1:t})$, is determined by the following three components:

- $$p(z_t|o_{1:t-1}) = \sum_i p(z_t|z_{t-1} = l_i)p(z_{t-1} = l_i|o_{1:t-1}), \tag{10}$$

  where l_i is the i-th activity, and $p(z_t|z_{t-1})$ is the transition probability for the first-order Markov chain model.
- $p(z_t|o_t)$: the current *a posterior* which can be obtained from the proposed feature fusion framework.
- $p(z_t)$: the prior distribution for different activities.

To get $p(z_t|o_{1:t-1})$ from Equation (10), we need to obtain the transition probability $p(z_t|z_{t-1})$ for the first-order Markov chain model. Here, we model human activity sequence as a Markov chain, which describes the transition from one activity to another. Given the n activities $\{l_1, l_2, ..., l_n\}$, the i-th row and j-th column entry of the transition probability matrix, $\mathbf{A} \in \mathbb{R}^{n\times n}$, can be expressed as

$$a^{ij} = p(z_t = l_i|z_{t-1} = l_j), i, j = 1, 2, ..., n. \tag{11}$$

We intend to calculate the transition probability matrix based on the training data. Given m steps human activity sequence, the transition probability from state l_j to state l_i, denoted as a_{ij} can be calculated as

$$a^{ij} = \frac{\sum_{t=2}^{m} \delta(z_t - l_i)\delta(z_{t-1} - l_j)}{\sum_{t=2}^{m} \delta(z_{t-1} - l_j)} \tag{12}$$

where

$$\delta(\alpha) = \begin{cases} 1 & \alpha = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Next, the probability $p(z_t|o_t)$ can be obtained from the proposed feature fusion framework. Since the last layer of the proposed feature fusion framework is a softmax layer, it will produce the probability for each activity based on inputs, i.e., current smartphone sensor measurements. Specifically, the current *a posterior* probability can be expressed as

$$p(z_t|o_t) = softmax(\mathbf{l}_t). \tag{13}$$

Finally, the probability $p(z_t)$ can be easily counted based on the training data as

$$p(z_t = l_i) = \frac{\sum_{t=1}^{m} \delta(z_t - l_i)}{m} \tag{14}$$

The implementation of the proposed MFAP for HAR is shown in Algorithm 1.

Algorithm 1 Proposed MFAP for HAR

Input: $\mathbf{A} = \{a^{ij}\}$, $\mathbf{b}_t = \{b_t^i\} = \{p(z_t = l_i|o_t)\}$, $\mathbf{c} = \{c^i\} = \{p(z_t = l_i)\}$, $i, j = 1, 2, ..., n$, $t = 1, 2, ..., T$.
Output: Full *a posterior*: $\mathbf{r}_t = \mathbf{p}(z_t|o_{1:t})$, predicted activity: $\mathbf{O}$

Initialisation: $t = 1$
$\mathbf{r}_1 = \{r_1^i\} = \mathbf{b}_1$
$O_1 = \arg\max_{l_i} \mathbf{r}_1$
Recursion
for $t = 2$ to T **do**
  for $i = 1$ to n **do**
    $r_t^i = \frac{b_t^i \sum_j a_{ij} r_{t-1}^j}{c^i}$ based on Equation (9).
  end for
  $O_t = \arg\max_{l_i} \mathbf{r}_t$
end for
return $\mathbf{O}$

V. Experiments

A. Data description

To evaluate the performance of the proposed approaches for HAR using smartphone sensors, we firstly use a public dataset from UCI [12]. A Samsung Galaxy SII smartphone which is attached to the waist of subjects with fixed orientation was used for data collection. Both three-dimensional acceleration and gyroscope data were collected. This dataset contains six

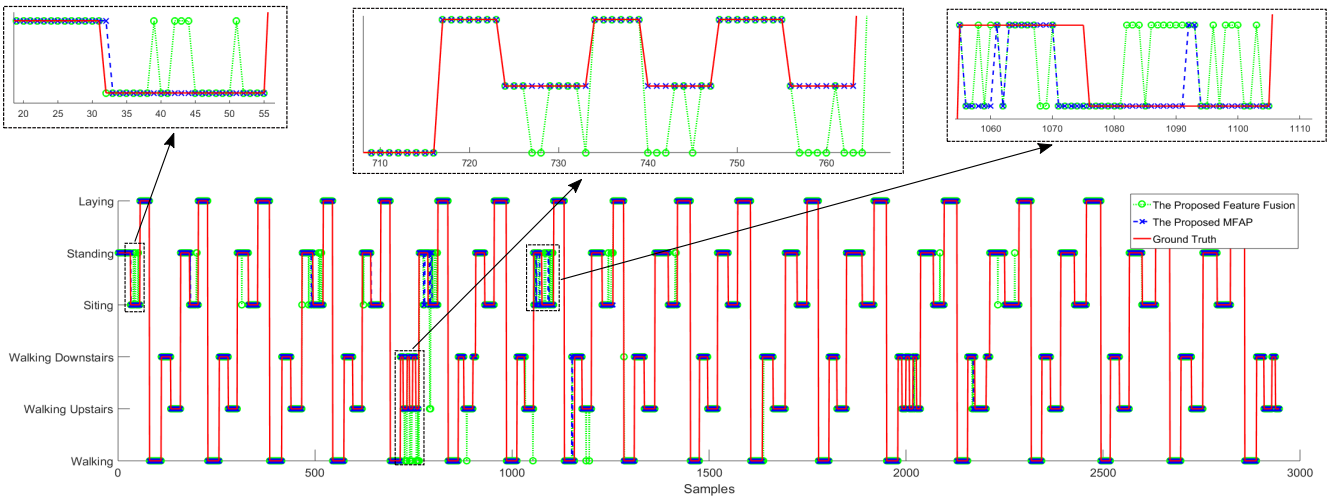
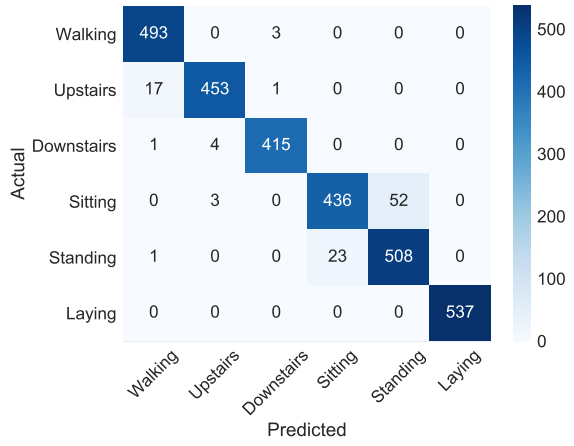
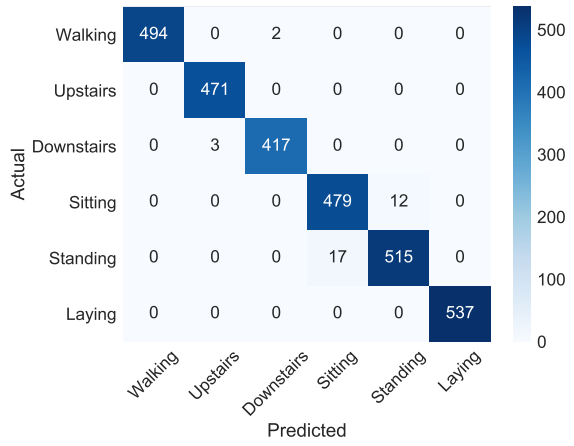


Fig. 3. The recognition results of the proposed feature fusion framework and the proposed MFAP on the public dataset.



(a) The proposed feature fusion framework



(b) The proposed MFAP

Fig. 4. The confusion matrices of the proposed feature fusion framework and the proposed MFAP on the public dataset.

activities, i.e., walking, walking upstairs, walking downstairs, standing, sitting and laying. The sampling frequency of the data is $50Hz$. A sliding window of 2.56 seconds (or a sample) with a 50% overlap is used for data segmentation. In total, 10299 samples are collected from thirty participants.

We also collected our own dataset using a recently released Huawei P20 Pro smartphone. For this dataset, instead of attaching the smartphone to a fix position which may not be realistic, we freely put the smartphone in three common positions, i.e., pants' pocket, shirt's pocket, and backpack, without any restrictions for data collection. Here, we consider some different activities, including walking, fast walking, running, walking upstairs, walking downstairs, and static. Similarly, we collected both three-dimensional acceleration and gyroscope with a sampling rate of $50Hz$. We also use a sliding window of 2.56 seconds with a 50% overlap for data segmentation. Totally, 4752 samples are collected from twelve volunteers.

For the public dataset and our own dataset, there are some differences: 1) The smartphones for experiments are different. 2) The placements of smartphones are different. 3) Due to the different smartphone placements, the explored activities in the two datasets are different. Since the smartphone is attached to the waist of subjects with fixed orientation in the public dataset, it is possible to detect the activities of "Standing" and "Sitting" based on the slight variances of smartphone orientations. Meanwhile, the orientation of "Laying" is totally different from the other two static activities of "Standing" and "Sitting", and thus various algorithms achieve very high recognition accuracy for "Laying" as shown in Table II. For these three activities, i.e., "Standing", "Sitting" and "Laying", in the public dataset, we can distinguish them based on the orientation information. However, for our own dataset, the smartphone is freely put in three common positions, without any restrictions on its orientation. Therefore, we are not able to distinguish the above three activities of "Standing", "Sitting" and "Laying" based on the orientation information. For this reason, we explore some other common activities, such as "Fast walking", "Running" and "Static" in our own dataset.

For both the public data and our own data, we random select 70% of the data to train different algorithms and the remaining for testing.

TABLE II
THE RECOGNITION ACCURACIES OF ALL THE APPROACHES ON THE PUBLIC DATASET

Method	Walking	Walking Upstairs	Walking Downstairs	Sitting	Standing	Laying	Overall
ANN	0.9899	0.9427	0.8262	0.8839	0.9586	0.9981	0.9372
ELM	0.9758	0.9512	0.8833	0.8615	0.9380	0.9963	0.9365
SVM	0.9919	0.9597	0.9000	0.8635	0.9173	1.0000	0.9403
RF	0.9698	0.9066	0.8405	0.8900	0.9248	1.0000	0.9253
Deep LSTM	0.9435	0.9766	0.9929	0.7434	0.9023	1.0000	0.9253
Proposed fusion	0.9940	0.9618	0.9881	0.8880	0.9549	1.0000	**0.9644**
Proposed MFAP	0.9960	1.0000	0.9929	0.9756	0.9680	1.0000	**0.9885**

TABLE III
THE RECOGNITION ACCURACIES OF ALL THE APPROACHES ON OUR OWN DATASET

Method	Walking	Fast Walking	Walking Upstairs	Walking Downstairs	Running	Static	Overall
ANN	0.9283	0.9839	0.9240	0.9283	0.9863	0.9915	0.9565
ELM	0.9578	0.9677	0.9480	0.9494	0.9863	0.9915	0.9663
SVM	0.9451	0.9758	0.9240	0.9114	0.9909	0.9957	0.9565
RF	0.9578	0.9919	0.9600	0.9536	0.9863	0.9957	0.9741
Deep LSTM	0.9826	0.9918	0.9536	0.9675	0.9913	0.9916	0.9797
Proposed fusion	0.9789	0.9839	0.9840	0.9831	0.9954	0.9957	**0.9867**
Proposed MFAP	0.9916	1.0000	0.9920	0.9958	1.0000	0.9957	**0.9958**

B. Experimental setup

To verify the performance of the proposed approaches, we compare with some advanced learning algorithms for HAR, including shallow learning algorithms with handcrafted features, such as artificial neural network (ANN), SVM [33], extreme learning machine (ELM) [34] and random forest (RF), and the deep learning algorithm of deep LSTM [35]. The parameters of all the benchmark approaches and the proposed approach are carefully tuned using a validation set. For ANN and ELM, the number of hidden nodes is determined by using grid search with the validation set. The popular radial basis function (RBF) kernel is chosen for SVM. The parameters of RBF kernel are determined using grid search. For RF, the number of decision trees is set as 500 for ensemble learning. The deep LSTM consists of two LSTM layers with sizes of 32 and 64, a FCL with a size of 100, and a softmax layer for classification. For the proposed fusion framework, two LSTM layers with sizes of 32 and 64 are used. The FCLs in Fig. 2 both have 100 hidden nodes.

C. Experimental results

1) Results on the public dataset: The experimental results on the public dataset are shown in Table II. With expert knowledge, conventional machine learning approaches of ANN, ELM and SVM with the handcrafted features slightly outperform the deep LSTM with automatic feature learning on the public dataset. This means that the handcrafted features are more representative for these activities. The proposed feature fusion framework which combines handcrafted features and automatically learned features by the deep algorithm has a superior performance over these benchmark shallow and deep algorithms. This indicates that handcrafted features and automatically learned features by the deep algorithm contain unique information for HAR and can complement each other, leading to a better performance. By taking the dynamics of human behavior into consideration, the proposed MFAP achieves the best performance. The overall accuracy is as high as 98.85%.

We now zoom into the performance of specific activity classification. Among all the activities, the activity of "Laying" has the highest recognition accuracy, due to the distinct smartphone orientation for this activity against these of the other five activities. The activities of "Sitting" and "Standing" have very similar patterns on smartphone sensor readings. Therefore, the recognition accuracies of these two activities are relatively low. Similarly, the recognition performances of the activities of "Walking Upstairs" and "Walking Downstairs" are also limited, because of the similar sensory patterns. Owing to the proposed feature fusion framework and the consideration of the dynamics of human behaviour, the proposed MFAP has the highest recognition accuracy for all the six activities.

We have shown the activity recognition results of the proposed feature fusion framework and the proposed MFAP for testing in Fig. 3. It can be observed that the activities of "Standing" and "Sitting" are difficult to separate, due to the similar sensory patterns. The activities of "Walking", "Walking Upstairs" and "Walking Downstairs" suffer from the same issue. By taking the dynamics of human behaviour, the proposed MFAP algorithm dramatically improves the results. This clearly indicates the effectiveness of the proposed MFAP algorithm for HAR. Fig. 4 shows the confusion matrices of the proposed feature fusion framework and the proposed MFAP on the public dataset. The general conclusion is the same. By considering human dynamics, the proposed MFAP improves the recognition accuracies for all the six activities.

2) Results on our own dataset: The experimental results on our own dataset are shown in Table III. Generally, all the approaches perform better on our own dataset when compared with the public dataset. One possible reason for the distinct results is that the explored activities are different for the two datasets. Based on Table II, we can find that the activities of "Standing" and "Sitting" are difficult to be separated, due to the similar sensory patterns (no movement

and similar smartphone orientation). While the activities in Table III are relatively easier to be separated. Moreover, the different devices for data collection and the way how the data was collected for the two datasets may also contribute.

Different from the results on the public dataset, the Deep LSTM with automatically learned features outperforms the conventional machine learning approaches with handcrafted features. This means that the automatically learned features by the deep algorithm are more representative for HAR on this dataset. Similarly, the proposed fusion framework which combines the handcrafted features and the automatically learned features by the deep algorithm outperforms the deep algorithm of deep LSTM and the conventional machine learning approaches, i.e., ANN, ELM, SVM and RF, with handcrafted features. We can conclude that the handcrafted features and the features learned by the deep algorithm have unique merits, resulting distinct performances on different datasets. With the proposed feature fusion framework, we can make good use of the merits of these two types of features to boost the performance for HAR using smartphone sensors. In addition, the proposed MFAP is able to take the dynamics of human behaviour into consideration, further improving the performance of the proposed feature fusion algorithm. The overall accuracy is as high as 99.58% on our own dataset.

For our own dataset, we consider some different activities due to the different placement of smartphones in the two datesets. It can be found that the activities of "Fast Walking", "Running" and "Static" which contain distinct movement patterns that can be easily identified with high recognition accuracies. However, activities of "Walking", "Walking Upstairs" and "Walking Downstairs" have very similar movement patterns, and thus confuse most of algorithms. Owing to the proposed feature fusion framework and the consideration of the dynamics of human behaviour, the final recognition accuracies of the proposed MFAP are higher than 99% for all the activities.

Fig. 5 shows the recognition results of the proposed feature fusion framework and the proposed MFAP for testing on our own dataset. Even though the proposed feature fusion framework has already achieved a very high recognition accuracy, i.e., 98.67%, it still contains some wrong estimations, shown as many spikes (see green line in Fig. 5) which are harmful for real applications, such as home automation. With the proposed MFAP which takes the dynamics of human behaviour into consideration, most of the wrong estimations can be corrected. We also show the confusion matrices of the proposed feature fusion framework and the proposed MFAP on our own dataset in Fig. 6. It can be found that the proposed MFAP corrects most of the wrong predictions of the proposed feature fusion framework, owing to the consideration of human dynamics.

3) Compared with state-of-the-arts: We have also compared with some state-of-the-art approaches in the literature, including HNN [15], FW KNN [14], FW Naive Bayes [14], HF-SVM [33], Two-stage CHMM [16], SRC-SVD [17], FMM-CART [18], SAEs-c [19], Convnet [20], HCF Convnet [21], tFFT Convnet [20] and Knowledge Distilling [13], using the public dataset. The detailed reviews of all these approaches can be found in Section II. TABLE IV demonstrates the

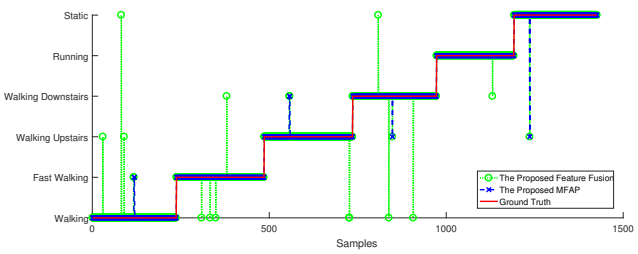
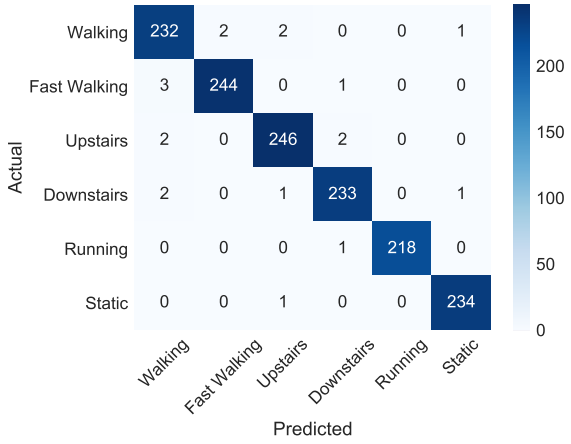
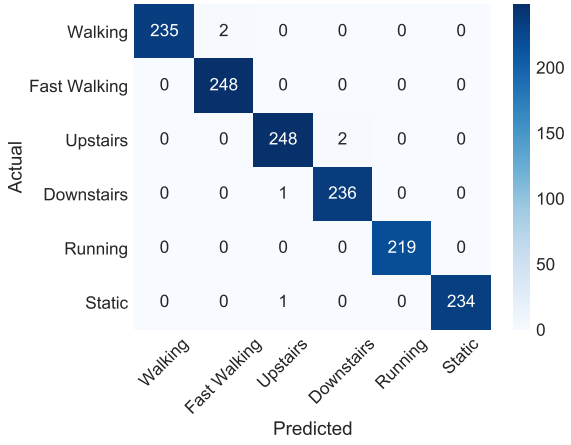


Fig. 5. The recognition results of the proposed feature fusion framework and the proposed MFAP on our own dataset.



(a) The proposed feature fusion framework



(b) The proposed MFAP

Fig. 6. The confusion matrices of the proposed feature fusion framework and the proposed MFAP on our own dataset.

experimental results of these state-of-the-arts and the proposed approach. It can be found that our proposed approach is able to achieve a superior performance over these state-of-the-art methods.

VI. Conclusion

In this paper, we firstly propose a feature fusion framework which combines handcrafted features with domain knowledge and automatically learned features by a deep algorithm, for

TABLE IV
COMPARISON WITH STATE-OF-THE-ARTS

Methods		Accuracy
Shallow algorithms	HNN [15]	87.4%
	FW KNN [14]	87.8%
	FW Naive Bayes [14]	90.1%
	HF-SVM [33]	89%
	Two-stage CHMM [16]	91.76%
	SRC-SVD [17]	95%
	FMM-CART [18]	96.52%
Deep algorithms	SAEs-c [19]	92.16%
	Convnet [20]	94.79%
	HCF Convnet [21]	95.75%
	tFFT Convnet [20]	95.75%
	Knowledge Distilling [13]	97.35%
Proposed MFAP		**98.85%**

human activity recognition (HAR). By taking the dynamics of human behaviour into consideration, we then formulate a maximum full *a posteriori* (MFAP) with the past information and the current *a posterior* information obtained from the proposed feature fusion framework to give an optimal estimation of human activities. We employ a public dataset and a self-collected dataset to evaluate the performance of the proposed approaches. Extensive experiments show the proposed feature fusion frameworks outperforms 5 benchmark approaches. And the proposed MFAP can further improve the performance for HAR. We also compared with some state-of-the-art methodologies on the public dataset. The proposed MFAP achieves the best performance, indicating our proposed method is practical to be applied for real-world applications.

In our future works, we intend to focus on the recognition of some more complex activities [36]. Moreover, considering the variation of smartphone orientation, the recognition performance may degrade. How to enhance the performance of smartphone based HAR with varying device orientations is one of our future works.

REFERENCES

[1] Y. Zhang, G. Tian, S. Zhang, and C. Li, "A knowledge-based approach for multiagent collaboration in smart home: From activity recognition to guidance service," *IEEE Transactions on Instrumentation and Measurement*, 2019.

[2] O. D. Lara, M. A. Labrador *et al.*, "A survey on human activity recognition using wearable sensors." *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[3] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1147–1153.

[4] G. Panahandeh, N. Mohammadiha, A. Leijon, and P. Hndel, "Continuous hidden markov model for pedestrian activity classification and gait analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 5, pp. 1073–1083, 2013.

[5] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE sensors journal*, vol. 15, no. 3, pp. 1321–1330, 2015.

[6] Z. Chen, Q. Zhu, Y. C. Soh, and L. Zhang, "Robust human activity recognition using smartphone sensors via ct-pca and online svm," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 3070–3080, 2017.

[7] Q. Zhu, Z. Chen, and Y. C. Soh, "A novel semi-supervised deep learning method for human activity recognition," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 3821–3830, 2019.

[8] Z. Chen, C. Jiang, and L. Xie, "A novel ensemble elm for human activity recognition using smartphone sensors," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 2691–2699, 2018.

[9] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition." in *IJCAI*, vol. 15, 2015, pp. 3995–4001.

[10] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, "Deep activity recognition models with triaxial accelerometers." in *AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments*, 2016.

[11] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," *arXiv preprint arXiv:1604.08880*, 2016.

[12] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones." in *ESANN*, 2013.

[13] Z. Chen, L. Zhang, Z. Cao, and J. Guo, "Distilling the knowledge from handcrafted features for human activity recognition," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4334–4342, 2018.

[14] A. Wang, G. Chen, J. Yang, S. Zhao, and C.-Y. Chang, "A comparative study on human activity recognition using inertial sensors in a smartphone," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4566–4578, 2016.

[15] M. Eastwood and C. Jayne, "Evaluation of hyperbox neural network learning for classification," *Neurocomputing*, vol. 133, pp. 249–257, 2014.

[16] C. A. Ronao and S.-B. Cho, "Human activity recognition using smartphone sensors with two-stage continuous hidden markov models," in *Natural Computation (ICNC), 2014 10th International Conference on*. IEEE, 2014, pp. 681–686.

[17] R. Rana, B. Kusy, J. Wall, and W. Hu, "Novel activity classification and occupancy estimation methods for intelligent hvac (heating, ventilation and air conditioning) systems," *Energy*, vol. 93, pp. 245–255, 2015.

[18] M. Seera, C. K. Loo, and C. P. Lim, "A hybrid fmm-cart model for human activity recognition," in *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 182–187.

[19] Y. Li, D. Shi, B. Ding, and D. Liu, "Unsupervised feature learning for human activity recognition using smartphone sensors," in *Mining Intelligence and Knowledge Exploration*. Springer, 2014, pp. 99–107.

[20] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Systems with Applications*, vol. 59, pp. 235–244, 2016.

[21] ——, "Deep convolutional neural networks for human activity recognition with smartphone sensors," in *International Conference on Neural Information Processing*. Springer, 2015, pp. 46–53.

[22] D. Tao, Y. Wen, and R. Hong, "Multicolumn bidirectional long short-term memory for mobile devices-based human activity recognition," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 1124–1134, 2016.

[23] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, 2018.

[24] H. Qian, S. J. Pan, and C. Miao, "Sensor-based activity recognition via learning from distributions," in *AAAI*, 2018.

[25] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?" *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[26] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[27] Y. Liu, L. Nie, L. Liu, and D. S. Rosenblum, "From action to activity: sensor-based activity recognition," *Neurocomputing*, vol. 181, pp. 108–115, 2016.

[28] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] G. E. Hinton, "Learning multiple layers of representation," *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[31] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, no. 2, 2012.

[32] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-markov model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 838–845.

[33] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International workshop on ambient assisted living*. Springer, 2012, pp. 216–223.

[34] Y. Chen, Z. Zhao, S. Wang, and Z. Chen, "Extreme learning machine-based device displacement free activity recognition model," *Soft Computing*, vol. 16, no. 9, pp. 1617–1625, 2012.
[35] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie *et al.*, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks." in *AAAI*, vol. 2, 2016, p. 8.
[36] L. Liu, L. Cheng, Y. Liu, Y. Jia, and D. S. Rosenblum, "Recognizing complex activities by a probabilistic interval-based model." in *AAAI*, vol. 30, 2016, pp. 1266–1272.

Zhenghua Chen received the B.Eng. degree in mechatronics engineering from University of Electronic Science and Technology of China, Chengdu, China, in 2011, and Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2017. Currently, he is a scientist at Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include data analytics in smart buildings, ubiquitous computing, internet of things, machine learning and deep learning.

Chaoyang Jiang received the B.E. degree in electrical engineering and automation from China University of Mining and Technology in 2009, the M.E. degree in control science and engineering from Harbin Institute of Technology in 2011, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2017. He is currently an associate professor in the School of Mechanical Engineering, Beijing Institute of Technology. His research interests include statistical signal processing, sparse sensing, machine learning, and information fusion.

Shili Xiang is currently a scientist and principle investigator in Data Analytics Department, Institute for Infocomm Research. She received her Ph.D. degree in Computer Science from National University of Singapore (NUS) and B.S. degree in Computer Science from University of Science and Technology of China (USTC). Her research interests include smart mobility, ubiquitous computing, data mining and machine learning.

Jie Ding received her B.Eng. degree in automation from Harbin Engineering University, China in 2012 and Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore in 2018. She is currently a Scientist in Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore. Her research interests include machine learning, pattern recognition, control & optimization, and complex networks.

Min Wu is currently a senior scientist in Data Analytics Department, Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. He received his Ph.D. degree in Computer Science from Nanyang Technological University (NTU), Singapore, in 2011 and B.S. degree in Computer Science from University of Science and Technology of China (USTC) in 2006. He received the best paper awards in InCoB 2016 and DASFAA 2015. He also won the IJCAI competition on repeated buyers prediction in 2015. His current research interests include machine learning, data mining and bioinformatics.

Xiaoli Li is currently a principal scientist at the Institute for Infocomm Research, A*STAR, Singapore. He also holds adjunct professor positions at Nanyang Technological University. His research interests include data mining, machine learning, AI, and bioinformatics. He has been serving as a (senior) PC member/workshop chair/session chair in leading data mining and AI related conferences (including KDD, ICDM, SDM, PKDD/ECML, WWW, IJCAI, AAAI, ACL and CIKM). Xiaoli has published more than 180 high quality papers and won numerous best paper/benchmark competition awards.