SensCapsNet: Deep Neural Network for Non-Obtrusive Sensing Based Human Activity Recognition

Article *in* IEEE Access · January 2020

DOI: 10.1109/ACCESS.2020.2991731

7 authors, including:

Nguyen Thai Son
Posts and Telecommunications Institute of Technology
4 PUBLICATIONS **112** CITATIONS
SEE PROFILE

Vu Hai
Hanoi University of Science and Technology
136 PUBLICATIONS **1,012** CITATIONS
SEE PROFILE

Thanh-Hai Tran
Hanoi University of Science and Technology
150 PUBLICATIONS **1,011** CITATIONS
SEE PROFILE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 27, 2020, accepted April 23, 2020. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.2991731

SensCapsNet: Deep Neural Network for Non-Obtrusive Sensing Based Human Activity Recognition

CUONG PHAM[1], SON NGUYEN-THAI[1], HUY TRAN-QUANG[3], SON TRAN[2], HAI VU[4], THANH-HAI TRAN[4,5], AND THI-LAN LE[4,5]

[1]Department of Computer Science, Posts and Telecommunications Institute of Technology, Hanoi 100000, Vietnam
[2]Department of Computer Science, New Mexico State University, Las Cruces, NM 88003, USA
[3]Vietnam Post and Telecommunication-Information Technology (VNPT), Hanoi 100000, Vietnam
[4]International Research Institute MICA, Hanoi University of Science and Technology, Hanoi 10000, Vietnam
[5]School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi 10000, Vietnam

Corresponding author: Thi-Lan Le (thi-lan.le@mica.edu.vn)

This research was funded by Vietnam Ministry of Science and Technology under grant number DTDLCN-16/18 ''Automated Respiration Symptoms monitoring and Abnormal Human Activity Detection Using the Internet of Things''.

ABSTRACT Recently, the recent advancement of deep learning with the capacity to perform automatic high-level feature extraction has achieved promising performance for sensor-based human activity recognition (HAR). Among different deep learning methods, Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) have been widely adopted. However, scalar outputs and pooling in CNN only allow to get the invariance but not the equivariance. The capsule networks (CapsNet) with the vector output and routing by agreement is able to capture the equivariance. In this paper, we propose a method for recognizing human activity from wearable sensors based on a capsule network named **SensCapsNet**. The architecture of **SensCapsNet** is designed to be suitable for spatial-temporal data coming from wearable sensors. Experimental results show that the proposed network outperforms CNN and LSTM methods. The performance of the proposed CapsNet architecture is assessed by altering dynamic routing between capsule layers. The proposed **SensCapsNet** yields improved accuracy values of 77.7% and 70.5% for 1 routing on two testing datasets in comparison with the baseline methods based on CNN and LSTM that yields the F1-score of 67.7% and 69.2% for the first dataset and 65.3% and 67.6% for the second dataset respectively. Moreover, even several human activity datasets are available, privacy invasion and obtrusive concerns have not been carefully taken in to consideration in dataset building. Toward to build a non-obstructive sensing based human activity recognition method, in this paper, a dataset named **19NonSens** is designed and collected from twelve subjects wearing e-Shoes and a smart watch to perform 19 activities under multiple contexts. This dataset will be made publicity available. Finally, thanks to the promising results obtained by the proposed method, we develop a life logging application which achieves a real-time computation and the accuracy rate greater than 80% for 5 common upper body activities.

INDEX TERMS Human activity recognition, capsule net, wearable sensors.

I. INTRODUCTION

Human activity recognition (HAR) using non-obtrusive sensing techniques has recently received great attention from both researchers and industry. With the rapid progress in semiconductor technology, low cost sensors (e.g. accelerometers and gyroscopes) with small size, light weight and low power consumption could be easily embedded hidden inside low obtrusive wearable devices (e.g. smart-phones, smart-watches and smart-shoes). These wearable devices have being used more and more popular in daily life. Based on the signals collected from embedded sensors, the human activities will be able to be segmented, analysed and recognized by learning signal patterns. The main goal of sensing-based HAR turns to exploit a robust classification method so that it can overcome challenges usually faced by traditional machine

The associate editor coordinating the review of this manuscript and approving it for publication was Charith Perara.

learning methods using one-dimensional temporal sequences (i.e. noise, fixed-length of sliding windows, temporal correlations between the collected signals). Particularly, in case of recognizing various types of activities in daily life, proposing efficient discriminated features is primarily required.

Recently, deep neural networks have made a great advance in many classification tasks. Ones have shown their feasibility for automatically extracting and representing features in a hierarchy from low-level to high-level abstractions. Deep neural networks avoid heuristic parameters of conventional hand-designed features as well as scale better for more complex behavior-recognition tasks. Recent surveys on the deep learning methods for sensor-based activity recognition have shown the superior results of deep learning methods in comparison with hand-desinged features-based methods for human activity recognition [1], [2]. Among different deep learning methods, Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) have been widely adopted. However, scalar outputs and pooling in CNN only allow to get the invariance but not the equivariance. The capsule networks (CapsNet) with the vector output and routing by agreement is able to capture the equivariance. In this paper, we propose a method for recognizing human activity from wearable sensors based on a capsule network named **SensCapsNet**. The architecture of **SensCapsNet** is designed to be suitable for spatial-temporal data coming from wearable sensors. Two main kinds of signal used in our work are accelerometer and gyroscope. Experimental results show that the proposed network outperforms CNN and LSTM methods.

Moreover, in the context of ubiquitous human activity recognition (or recognizing people activities in common lives), some common public datasets (e.g., [3], [4]) have not been constructed under different contexts as indoor or outdoor scenes. Beside, the role of wearable sensors versus their mounting's positions on human-body (e.g., watches for monitoring the activities of upper extremities, shoes for lower ones) have been not clearly analyzed. These reasons motivate us to construct a new dataset with various types of human activities in both outdoor and indoor scenes. Issues of mounting position are also taken into account for monitoring activities of both lower and upper extremities.

The main contributions of the paper are as follows:

- A new sensing-based HAR dataset (named **19NonSens**) is built. To collect the human activity, we use a commercial smart-watch (Samsung Gear G2) which is built-in sensors and our self-made smart-shoes embedded with tiny wireless accelerometers (named as e-Shoe) for data acquisition. This design allows maximizing unobtrusiveness to subjects as well as allows them to comfortably perform daily activities in a realistic manner. 19 activities including *null* activities have been designed and collected from 12 subjects in both indoor and outdoor scenes.
- A new method based on capsule network for human activity recognition (**SensCapsNet**) is proposed. The architecture of **SensCapsNet** is designed to be suitable for spatial-temporal data coming from werable sensors (e.g., accelerometers and gyrocopes).
- A real-time human activity recognition and logging application has been built to illustrate the potential applications of using non-obstructive sensing data for human activity recognition.

The remainder of this paper is organized as follows: Section II briefly reviews related works on sensing-based HAR. Section III presents in detail the **19NonSens** dataset. The proposed method based on capsule network is described in Sections IV. Section V reports comparative evaluations and Section VI describes the application for human activity recognition and logging. Finally, discussions and conclusions are presented in Section VII.

II. RELATED WORK

Human activity recognition (HAR) based on wearable sensors has been intensively attempted in the literature. Readers can refer comprehensive surveys related to this topic in [1]–[3]. At the heart of a wide range of practical applications [2], [5], HAR basing on wearable sensors offer assistive technologies for healthcare [6]–[8], and helping the elderly or people with special health conditions (e.g., dementia) to live more independently at their homes. For instance, [9]–[11] proposed solutions for healthier cooking. Works in [12], [13] offered intelligent homes. In this section, we briefly review works aiming at tackling two major issues usually raised when deploying feasible applications. First, we survey HAR's works which attempt to use non-invasive or unobtrusive sensing. Second, we review advanced techniques for wearable-based HAR and their evaluation on benchmark datasets. Particularly, the works utilizing the recent Deep Neural Networks (DNN), will be described.

A. UNOBTRUSIVE TECHNOLOGIES FOR HAR

Pervasive or unobtrusive sensing based activity recognition could be understood as the technologies ensure invisibility to the users by embedding sensors into the subject as natural as possible. Work by Pham *et al.* [10], for instance, deployed multiple accelerometers inside kitchen appliances for detection of fine-grained cooking activities such as chopping, scooping, stirring, etc. The application presented in [10] is to help dementia people to live more independently at their homes. Similarly, high-level activities such as making cereal and coffee have been addressed by Buettner *et al.* [11]. The authors attached Radio Frequency Identification (RFID) tags on food containers such as the jug and bowl for recognition of activities by inferring objects getting involved a specific cooking task. A work proposed by Tapia, M. *et al.* [14] employs numerous simple and binary sensors at a smart home for detection of in-home activities such as bathing, cleaning, etc. Recent work such as [15] employed RF sensors mounted under the work surface for recognizing clerk and desk-work activities under real office settings. In the above

works, the wearable devices are completely integrated into the environment and are invisible from the users. Therefore, they allow the users to perform their activities in a non-invasive and unobtrusive manner.

Recently, the use of smart devices such as smart-phone, smart-watch has been more and more popular. The inertial sensors (e.g., accelerometer, gyroscope) are usually de-facto built-in inside such devices. This offers solutions to maximize the unobtrusive manner, particularly, in the context of monitoring human daily activities. There are a number of works investigating advantages of smart-phone, smart-watch, smart-shoe for detecting human activities and/or mobility. For example, Kwapisz *et al.* [16] exploit advantages of smart-phone for recognizing walking, jogging, standing, climbing stairs, and sitting. Several features such as the average time between peaks, standard deviation, bin distribution are manually extracted from sensing data streams. These features then are utilized to train and test classification methods including decision tree, logistic regression, multilayer perceptron. Similarly, work by Xing *et al.* [17] detects several mobility activities. Some works exploit features extracted from accelerometer built-in smart-watch. In [18], the authors detect drinking activity with over 93% accuracy. Other works attempt non-invasive activity recognition by embedding and hiding the sensors inside the fabrics. Such devices can be worn by human such as shoe [19], [20] or textile [21]. However, classifying activities performed with both hands and foots such as drinking, brushing, running, walking seems to be a considerable challenge for smart-device-based activity recognition. In this study, we deploy both two smart-watches and a smart-shoe human worn on human body to address this issue. This sensor mounting allows users to comfortably perform their daily activities under realistic settings. Particularly, a set of 19 activities covering both upper and lower extremities are collected in both indoor and outdoor environments.

B. METHODS FOR HAR USING WEARABLE SENSING

Many HAR methods focus on recognizing everyday activities such as running, biking, walking, cooking, walking, jogging, standing, climbing stairs, sitting, or even fine-grained cooking activities such as chopping, scooping, stirring etc. In early works, intrinsic temporal sequences in human activities have been processed by implementing hidden Markov models (HMMs) above the RBM layers. A series of related techniques have been listed in a survey of Lara et al [2]. It is worth to mention that there are some limitations or obstacles from current techniques: it is not easy to capture several daily activities such as preparing food or cleaning house using a small mobile device such as a smart phone; detecting fine-gained activities performed with hands such as drinking or brushing. Furthermore, achieving the trade-off between sensor sampling frequency and recognition accuracy for real-time implementation on a smart-phone is a challenge. Therefore, it still remains a considerable challenge for smart-wearable-based activity recognition.

Most approaches to HAR using wearable sensors focus on recognizing a pre-defined set of activities [2]. However, recognizing *null* activities (arbitrary out of interest activities) or recognizing a larger set of activities in different contexts using multiple sensors but hidden from the users needs to resolve imbalance classification. In [22], the FE-AT (Feature-based and Attribute-based learning) approach has been proposed to address this issue. FE-AT focuses on the shortage of labeled data by leveraging the relationship between existing and new activities. Recently, the use of convolutional neural networks (CNNs) for HAR was introduced in [23]. The authors deployed a simple CNN model for learning and recognizing data from single accelerometer. Another model in [24] used deep CNNs in a multi-sensor recognition framework which built a new multi-channel time series architecture of CNNs. The architecture proposed in [25] uses deep recurrent neural networks (DRNNs) for building recognition models that are capable of capturing long-range dependencies in variable-length input sequences. In their work, effectiveness of long short-term memory (LSTM) in DRNNs is confirmed on miscellaneous benchmark datasets.

CapsNets were fisrt introduced in 2017 for image classification task and has obtained superior performance on the MNIST dataset in comparison with the state of the art CNN-based methods [35]. Since then, there has been an upsurge in employing Capsule Networks for different computer science tasks. Recently, previous studies have tried to extend CapsNet for working with temporal information such as bearing fault diagnosis on raw vibration signals [26] or continuous sign language recognition from wearable IMUs [27]. However, to the best of our knowledge, this is the first work where capsules are employed for sensor-based activity recognition.

III. 19NONSens - NON-OBSTRUCTIVE SENSING HUMAN ACTIVITY DATASET

A. HARDWARE SETUP

To collect human activity dataset, we use two devices that are a Samsung Gear G2[1] (SG2) and a self-made smart-shoes embeddedwith tiny wireless accelerometers (named as e-Shoe). Figure 1 shows these devices images as well as wearing positions. The Smart-watch SG2 employs different sensors such as an accelerometer, a gyroscope, a heart rate sensor, a themal and a light sensor. In this study, sensing signals from the accelerometer and the gyroscope will be used as inputs of the system. For simpler synchronization, both accelerometer and gyroscope sensors of SG2 are set to the sampling frequency of 50Hz which is identical to the sampling rate of the 3-axis wireless accelerometers (WAX3). SG2 will be worn on body's hand as shown in Fig. 1(b).

Instrumented inside e-Shoe, WAX3 (Fig. 1(c)) is a MEMS accelerometer developed by researchers at Open Lab [9]. Figure 1(d) shows the position of a WAX3 sensor is embedded and hidden into the sole of e-Shoe. WAX3 weights 7 grams and dimensions of 23 x 32.5 x 7.6 mm, and operates with

[1] https://www.samsung.com/global/galaxy/gear-s2/

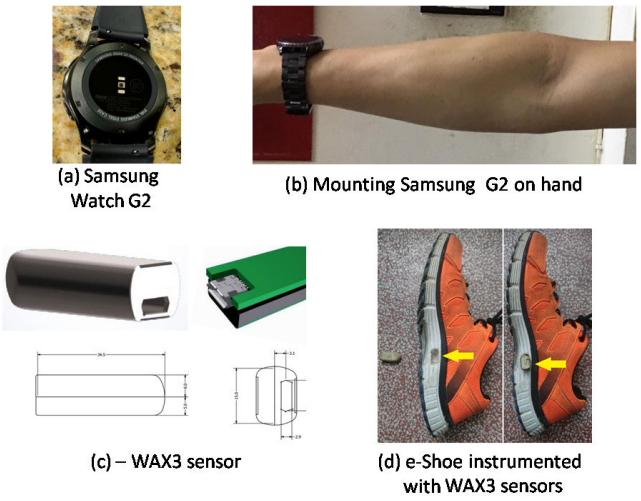


FIGURE 1. Setup smart-watch (upper row) and smart-shoe (lower row) as unobtrusive wearable sensors.

the IEEE 802.15.4 low power radio protocol. It can perform a sampling rate up to 2ks.sec-1 and be adaptable to Open Sound Control (OSC) message, binary, and American Standard Code for Information Interchange (ASCII) formats. There are two WAX's versions: WAX receiver connects to the computer via a USB port and WAX transmitter can wirelessly communicate to the receiver with the sensing range up to 25 meters. WAX is also equipped a re-chargeable Li-Polymer battery with the battery life is up to 8 hours for continuously transmitting signals and up to 56 days for hibernate mode. The acceleration signals of the WAX sensor embedded inside the shoes are measured in X, Y, and Z axes (relative to the accelerometer) and three directions of the movement (X, Y, Z) can be computed through tilt angles. Acceleration values are transmitted with a sampling frequency of 50Hz (50 samples per second). To ensure the sensor can be chargeable, the antenna of WAX points inside the shoes while the (female) hole towards outside WAX3 is easy to be embedded and hid inside the insole of e-Shoe.

B. DATASET CONSTRUCTION

We collected and annotated data in indoor and outdoor contexts to build up a dataset for experimental evaluation. The constructed dataset comprises of 18 activities plus *Null* activities. The list of activities and roles of each type of sensor in each action is given in Tab.1.

Twelve subjects aged between 19 and 45 are asked to worn e-Shoes and Samsung Gear S2 smart-watch on the preferred hand (10 right-handed and 2 left-handed). The subjects are asked to sign the consent forms and given the list of 18 activities. Before performing activities, subjects are asked to perform the "kick and hit-the-hand" activity to make highly distinctive signals for synchronizing sensors and video, and then resting for 10 seconds before performing activities (see Fig. 3. a). There are 9 in-door activities such as brushing, slicing and 9 sport out-door activities such as kicking, running. During performing the pre-defined activities, the subject could perform any arbitrary activity out of 18 activity list. We consider all of activities out of interest as *Null* activities. Duration time for each activity varies from 3 to 10 minutes. In addition, several surveillance cameras are installed in the kitchen, living room, and outdoor space to capture the activity videos which are used later for annotation (see Fig. 2). Two people have annotated the whole dataset using ELAN software.[2] Only signal corresponds to predefined activities are labeled and the other are marked as *Null*. Figure 3. b-d show some examples of synchronized signal of brushing, peeling and kicking in **19NonSens** dataset.

TABLE 1. List of human activities captured with Smartwatch and e-Shoes in indoor (i) and outdoor (o) environments. The activities are mostly performed by upper (u) or and lower (l) human body part.

No	Activity label	in/out env.	upper/lower movement	Duration (minutes)
1	Brushing	i	u	78
2	Washing hand	i	u	48
3	Slicing	i	u	71
4	Peeling	i	u	72
5	Up-stair	i	u+l	58
6	Down-stair	i	u+l	56
7	Mixing	i	u+l	52
8	Wiping	i	u+l	63
9	Sweeping floor	i	u+l	102
10	Turning shoulder	o	u	46
11	Turning wrist	o	u	51
12	Turning knee	o	l	53
13	Turning haunch	o	l	52
14	Turning ankle	o	u+l	49
15	Walking	o	u+l	64
16	Kicking	o	u+l	77
17	Running	o	u+l	61
18	Cycling	o	l	76
19	Null	i/o	u/l	178

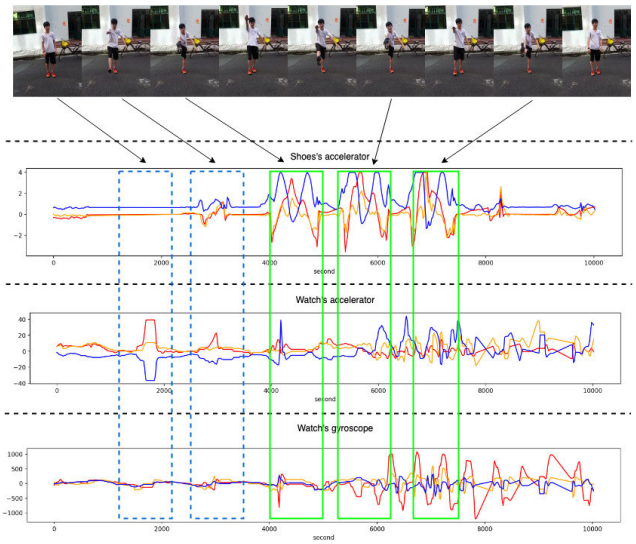
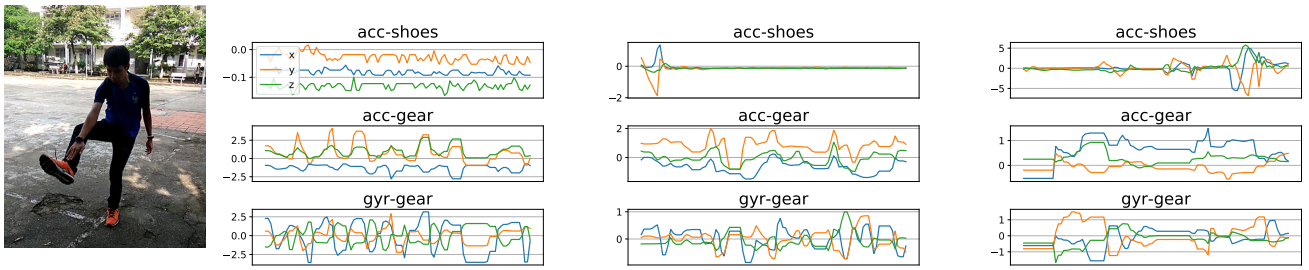


FIGURE 2. Signal synchronization in 19NonSens dataset.

[2]https://tla.mpi.nl/tools/tla-tools/elan/



(a) Kick and hit-the-hand (b) Tooth brushing (c) Peeling (d) Kicking

FIGURE 3. **(a) Synchronization trigger and (b, c, d) examples of activities of 19NonSens (acc-shoes: accelerometer values of e-Shoes; acc-gear: accelerometer values of SG2; gyr-gear: gyroscope sensing values of SG2).**

TABLE 2. **Comparison of the 19NonSens dataset with recent benchmark public datasets (as listed in [25]).**

Datasets	#Act	#Sub	Body part	Sensor	Obtr.
UCI-HAD (29)	6	30	Lower	Acc., Gyro	Low
USC-HAD (3)	12	14	Lower	Acc., Gyro	Low
Opportunity (30)	18	4	Upper	IMU, Acc., Gyro	High
FOG (31)	2	10	Upper	Acc.	Low
Skoda (32)	11	1	Upper	Acc.	Low
19NonSens	19	12	Upper, Lower	Acc., Gyro	Low

This dataset will be publicly made accessible through our github page.[3] Comparison of the constructed dataset with the most recent benchmark datasets of HAR using wearable sensors [25] is presented in Tab.2. As reported, the constructed dataset has some advances and challenges. First, it covers both indoor and outdoor contexts. Second, the activities of both upper extremities (hands) and lower ones (foots) are attempted with the obstructiveness is miniaturized. Finally, the dataset is constructed by a reasonable number of participants who perform their activities as natural as possible without any instruction from the experimenters.

IV. SensCapsNet - CapsuleNet FOR HUMAN ACTIVITY RECOGNITION FROM ACCELEROMETER

A. PRE-PROCESSING TECHNIQUES

As signal from accelerometer-based sensing devices may contain noise due to the individual and environmental variations, sensor diversity and sensor placement issues, before feeding this signal to the network, we first apply some preprocessing techniques. Low-pass and high-pass filtering are applied for noise removals. In addition, as sometimes sensor signals can be dropped, we keep 2-second frames contains more than 70% of its full complement for next step, and discard on the grounds frames less than 70% as they are insufficient information to classify activities. After that, a cubic spline interpolation method is applied for re-sampling data to fill out the dropped samples. Then, sensing values are normalized into the range of $[-1,1]$. In our work, the sampling frequencies are set to 50Hz. This means, for each second we have 50 samples of X, Y, Z acceleration values. We then segment signal by using 2-second sliding windows with 30% overlapping between two consecutive sliding windows. The 2-seconds window with overlapping ration are inspired by the study in [19] as this would cover most of the activities of interest while reducing time delayed for real-time implementation.

[3]We will provide the URL upon the request

B. 1D-CONVOLUTIONAL OPERATOR

Since our data is time-dependent, we employ 1-dimensional convolution operation (1D-Conv) to extract local pattern. Assume the input feature is $a^{l-1} \in \mathbb{R}^{L\times D}$ where L is the number of time points in a frame, D is the size of feature set. The output of the 1D-Conv is presented in Equation 1:

$$a_i^{l,c} = b^c + \sum_{v=1}^{D}\sum_{u=1}^{k} w_{uv}^{l,c} a_{i-\frac{k}{2}+u,v}^{l-1} \qquad \forall c = 1,..,C \qquad (1)$$

where b^c is the bias term of the $c-th$ output feature in the set of C output features. k is the size of kernel which slices along the times axis, $w^{l,c}$ is the weight matrix at layer l regarding the c-th output feature.

C. SensCapsNet FOR HUMAN ACTIVITY RECOGNITION

Capsule network (CapsNet) was first introduced for image classification task [35]. A capsule is a group of neurons with can model different entities or parts of entities in one image. The capsules in a network will be undergo a routing by agreement algorithm which allow the network to capture the parts-to-whole relationship between entities and to learn viewpoint invariant representations. Recently, CapsNet has been applied to recognize events/activities from time series data such as traffic flow data [32] and video [33].

There are two main concepts in CapsNet that are *capsule* and *dynamic routing algorithms* between capsules. A capsule is a group of local neurons that encode the information into a vector using a complex internal computational process [34].

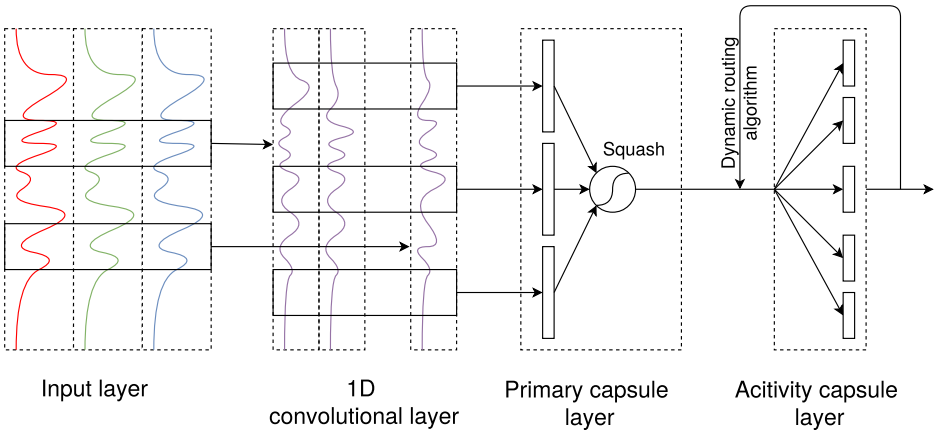


FIGURE 4. Architecture of the proposed SensCapsNet for human activity recognition.

A capsule is a combined series but not limited to convolutional layers, activation layers and fully-connected layers. Each capsule responses to an implicit pattern which is a restricted space. In image processing, that space can be a pose at different angles and sides [35].

In a conventional feed-forward network, the information from lower layer is passed to upper layer based on an unchanged learned set of parameters. However, in a capsule network, information is passed partially to highly agreeing capsules in the higher layer using dynamic routing algorithm. This algorithm dynamically modifies the weights of connection based on the agreement of the output and the input. In other words, it tries to identify the upper capsules which response to the claimed data and passes the information to them. The first step of this algorithm is to calculate a temporary output using a fair weight set for all connections. Then it estimates the responsiveness of a capsule by the similarity of the input and the output. Finally, the weights of connections are updated based on this analogy. A modification of connection weight is called a *routing iteration*. Each routing iteration changes the shares of information from capsule in the lower layer to the capsules in the upper layer.

Moreover, capsule network [35] introduced the *squash* function which scales a vector into another parallel vector whose length represents the probability of object presence while orientation represents the pose of object. Squash function is presented in Eq. 2 where vector s_j is scaled into v_j.

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \tag{2}$$

In this study, we propose a CapsNet with three-stage architecture for human activity recognition from wearable sensors. The network consists of a convolutional stage, a primary capsule stage and an activity capsule stage. The architecture of the network is illustrated in Fig. 4. The convolutional stage contains multiple 1D convolutional layers as presented in Eq. 1 and projection layers with ReLU activation function. This stage extracts abstract features for primary capsules from sensing data features. The primary capsule layer contains a large number of capsules. Each capsule encodes information into 8-dimensional vectors using a 1D convolution with a novel squash activation function [35]. The activity capsule layer contains as many numbers of capsules as the number of activities. The capsules in this stage connect densely to the capsules in primary capsule stage. Table 3 indicates the parameters of **SensCapsNet** for recognizing 19 activities in **19NonSens** dataset. The input is 9×100 dimensional vector with 9 is the X, Y, Z values of two accelerometers (one from the smart watch and one from the e-Shoe) and one gyroscope of the smart watch, 100 is the number of signal in 2 second window.

TABLE 3. Parameters of the proposed **SensCapsNet** for 19NonSens dataset.

Layer		Kernel size	Strides	Output dims
Conv	1D Conv1	5	1	256×96
Primary Caps	1D Cov2	5	2	256×46
	Squash	-	-	8×1472
ActivityCaps	AcitivyCaps	-	-	16×19

TABLE 4. Performance (%) on 19NonSens dataset.

	Precision	Recall	F1-score
CNN	67.5 ± 1.4	68.6 ± 1.4	67.7 ± 1.4
DeepConvLSTM	69.7 ± 1.1	69.4 ± 1.4	69.2 ± 1.3
SensCapsNet-1	$\mathbf{78.0 \pm 1.2}$	$\mathbf{78.8 \pm 0.9}$	$\mathbf{77.7 \pm 1.0}$
SensCapsNet-2	74.4 ± 1.0	74.9 ± 0.9	74.0 ± 1.0
SensCapsNet-3	72.4 ± 1.0	72.6 ± 0.8	72.0 ± 0.9

TABLE 5. Performance (%) on opportunity dataset.

	Precision	Recall	F1-score
CNN	66.1 ± 1.8	65.0 ± 1.7	65.3 ± 1.7
DeepConvLSTM	68.1 ± 1.7	67.3 ± 1.4	67.6 ± 1.4
SensCapsNet-1	$\mathbf{71.6 \pm 1.1}$	$\mathbf{69.9 \pm 1.3}$	$\mathbf{70.5 \pm 1.0}$
SensCapsNet-2	71.0 ± 1.6	69.1 ± 1.7	69.8 ± 1.5
SensCapsNet-3	70.8 ± 0.9	68.8 ± 1.5	69.6 ± 1.3

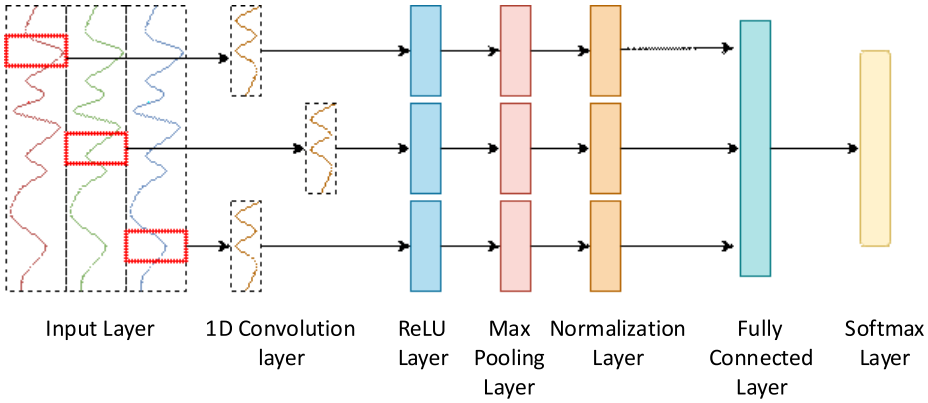


FIGURE 5. Architecture of the CNN used in our experiments for human activity recognition.

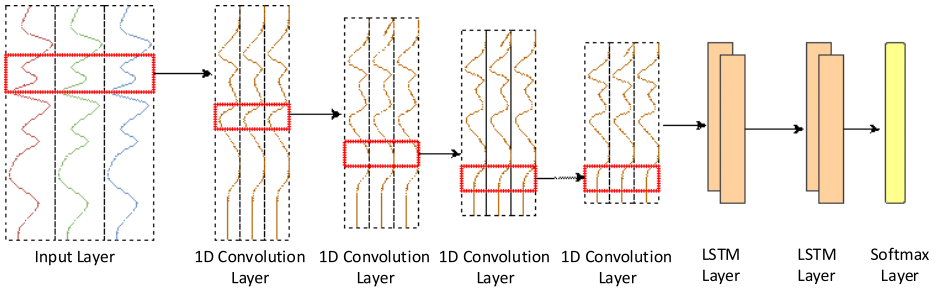


FIGURE 6. Architecture of the DeepConvLSTM used in our experiments.

To train this model, we use the length of output vector v at activity capsule stage to compute loss function presented in equation 3 where $m^+ = 0.9$; $m^- = 0.1$; $T_k = 1$ if the object k appears; the term λ represents the down-weighting of the loss for non-presented classes. In order to make the model, a simple reconstruction network as a regularizer in which β is the reconstruction regularization term; f is the reconstruction process and R is the square error between reconstructed data and input. This regularizer network contains three wide convolutional layers connected by two ReLU layers and a sigmoid function on the top to output auto-encoded data.

$$\begin{aligned} L = & \sum_{k=1}^{n} T_k max(0, m^+ - \|v\|)^2 \\ & + (1 - T_k)\lambda max(0, \|v\| - m^-)^2 \\ & + \beta R(f(x), x) \end{aligned} \tag{3}$$

In addition, inspired by the previous study [35], we apply reconstruction as regularization and set the regularization term β to 0.0001. As performance of the capsule network depends on the number of dynamic routing iterations. In our works, we report the performance of **SensCapsNet** with three iterations named **SensCapsNet-1**, **SensCapsNet-2** and **SensCapsNet-3** respectively.

D. CNN AND LSTM FOR ACTIVITY RECOGNITION

To evaluate the effectiveness of the proposed method, we will compare its performance with the baseline methods on common datasets. Two baseline methods are chosen in this paper. The first baseline to be investigated is the convolutional neural network CNN presented in [23]. This is a deep model allowing multichannel time series as inputs. This model consists of three stages. The first stages contains three modules, each of them works on the stream of a sensor. A module is a stack of four sets of four layers: a convolution layer, a rectified linear unit (ReLU) layer, a max pooling layer and a normalization layer. The second stage unifies the data of the three above streams using a fully connected layer that creates a parametric-concatenation. The final stage is a fully connected layer that maps the information into classes. Fig. 5 illustrates the architecture of the CNN model for human activity recognition.

The second baseline architecture used in this study is a deep model that combines both CNN and LSTM [24]. Although CNN is more sensitive than RNN in learning spatial relations from data, it is not designed for modeling long-term dependencies. On the contrary, LSTM [36] with three gates mechanism maintains the memory for an arbitrary number of computational steps. Therefore, a combination of CNN and LSTM is likely to be better in both recognizing local patterns and long relations. The DeepConvLSTM shown successfully evaluations on a series of benchmark datasets, as given in [25]. In this study we deploy a DeepConvLSTM which has four convolutional layers stacked on top of the raw sensor channel, as shown in Fig. 6. Those layers with convolution

TABLE 6. Detailed performance (%) of SensCapsNet-1 on 19NonSens dataset.

Activity	Precision	Recall	F1-score
Brushing	80.82	85.36	83.03
Hand washing	90.32	91.25	90.78
Slicing	91.16	95.91	93.47
Peeling	81.77	39.24	53.03
Up-stair	71.26	62.35	66.51
Down-stair	75.32	74.86	75.09
Mixing	81.48	89.55	85.32
Wiping	85.12	89.04	87.04
Sweeping floor	78.30	88.01	82.87
Turning shoulder	83.31	90.74	86.87
Turning wrist	81.39	88.16	84.64
Turning knee	77.32	89.18	82.83
Turning haunch	86.75	92.91	89.72
Turning ankle	50.74	44.65	47.50
Walking	73.09	85.32	78.73
Kicking	88.29	94.58	91.33
Running	71.80	69.00	70.37
Cycling	76.88	84.65	80.58
Null	56.52	42.39	48.45

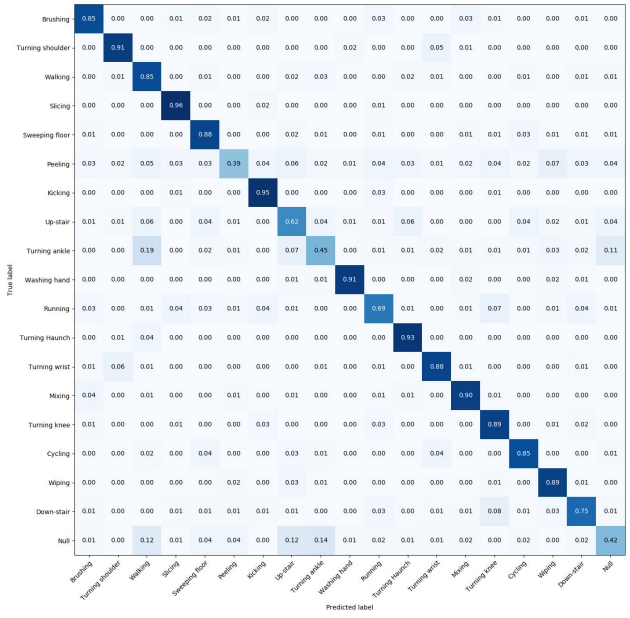


FIGURE 7. Confusion matrix of SensCapsNet-1 using both Smartwatch and e-Shoes sensors.

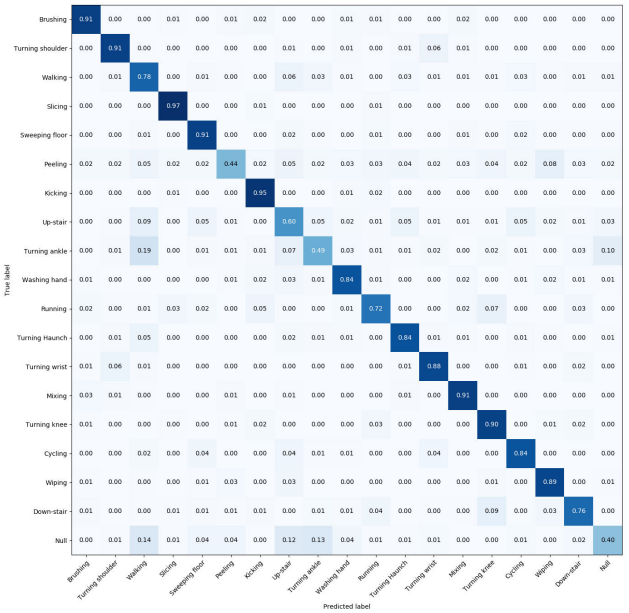


FIGURE 8. Confusion matrix of the SensCapsNet-1 on smart watch signals.

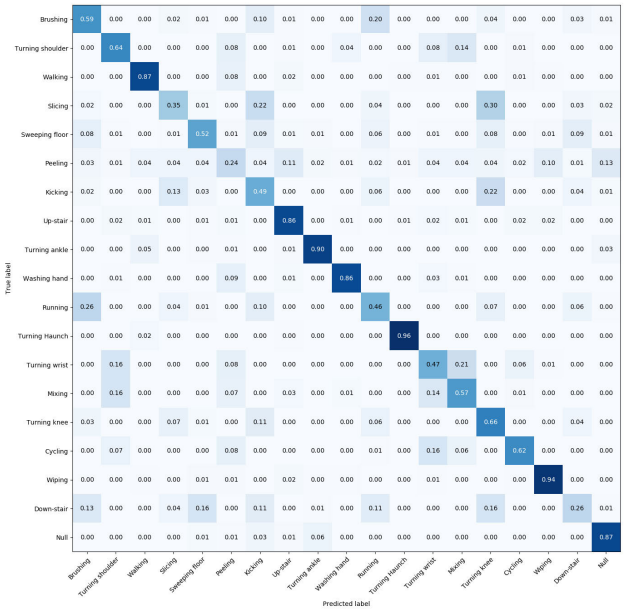


FIGURE 9. Confusion matrix of the SensCapsNet-1 on e-Shoes signals.

operations extract features for stacked LSTM layers. Following the study in [24], we stack two LSTM layers to enable the ability of modeling high level of abstraction.

V. EXPERIMENTAL RESULTS

A. DATASETS AND EVALUATION SETTINGS

1) EVALUATION DATASETS

There are two datasets used in our experiments: our **19NonSens** and **Opportunity** [29], [37]. Opportunity dataset provides signal of varied types of human activities including: periodic activities (e.g. walking), static activities (e.g. standing, lying down) and sporadic activities (e.g. opening a drawer). The activities in the dataset are hierarchically categorized into four abstract levels from atomic gestures such as moving bread to long sequential activities such as preparing breakfast. It comprises a very rich set of signals from various sensors mounted at different positions on human body. However, many of sensors and their positions violate the non-invasiveness and non-obstructiveness properties of the activity recognition task.

To ensure the obtrusiveness characteristics to end-users, we extract a subset from the **Opportunity** dataset that contains only the activities captured from three sensors attached on wrist and knee, including accelerometer channels (RKN_ and RLA) and 3D gyroscope channels (RLA). From which, the HL_Activity set of 6 labels *null* signal from three sensors that best fits our motivation and similar to those we collected in term of sensor types and sensors positioned on the human body while likely alleviating obstructiveness. They are the 3D accelerometer channels (RKN_ and RLA) and 3D gyroscope channels (RLA). We target at HL_Activity set of 6 labels including *null*. Our work differs from other works such as [25] used whole the Opportunity dataset and often ignored obtrusiveness characteristics.

2) EVALUATION SETTINGS

We employ the same protocol for both datasets: 10-fold cross validation. Under this protocol, the dataset is partitioned into 10 parts (folds), in which 9 parts are used for training, and the remaining one is used for testing, and the process is repeated for all parts, and the results are averaged after all. Three evaluation metrics that are Precision, Recall and F1-score.

B. EXPERIMENTAL RESULTS AND DISCUSSIONS

1) EXPERIMENTAL RESULTS ON **19NonSens** DATASET

Tab. 4 shows Precision, Recall, and F1-score achieved by different networks on our proposed **19NonSens** dataset. The recognition rates of the baseline models (CNN and DeepConvLSTM) are 68.6% and 69.4% respectively. The CNN model has lowest performance with precision of 67.5%, recall of 68.6%, and f1-score of 67.7%. DeepConvLSTM model, known as the original model designed for fusing multiple sensors, more effective than standard CNN [24]. On the proposed dataset, DeepConvLSTM model has slightly higher performance of around over 69%, which improve about 2% compared to the CNN model. This can be explained that our dataset is relatively complex as it covers various activities under different contexts which might challenge even deep models.

The proposed **SensCapsNet** significantly improves the recognition performance compared to the two baselines. Among three variations of Capsule network, the **SensCapsNet-1** with one routing iteration outperforms other capsule network variants. It achieves the highest recognition rates with both Precision and Recall of over 78%, and nearly 78% F1-score. Comparing to **SensCapsNet-2** and **SensCapsNet-3**, the increase of performance is about 4% and 6% respectively.

2) EXPERIMENTAL RESULTS ON **Opportunity** DATASET

The performance of network models on **Opportunity** dataset is shown in Tab. 5. The two baseline models achieved 65% to 67.3% in term of recall. DeepConvLSTM model is lightly 2% better than the standard CNN model, which is reasonable compared to [25] as we just used significantly less sensors than [25] (3 sensors in this study vs. 12 sensors in [25]). One again, the best performance (71.6% Precision, 69.9% Recall, and 70.5% F1-score) has been achieved by **SensCapsNet-1**, followed by its variants **SensCapsNet-2** and **SensCapsNet-3**. It proves the efficiency of capsule networks which takes information about the relative relationships between features into account. In the following, we will analyze in more detail the performance of **SensCapsNet-1** for each of activities and each of sensors on the **19NonSens** dataset.

3) DETAILED ANALYSIS OF **SensCapsNet-1**'s PERFORMANCE ON **19NonSens** DATASET

Table 6 shows performance obtained by **SensCapsNet-1** on **19NonSens** dataset. As can be seen, three most distinguishable activities are *Slicing*, *Hand washing* and *Kicking* with F1-score over 90%. The highest recall on *Slicing* activity is 95.91% which is a very promising result. This is explained by the fact that both accelerometer and gyroscope data represent rotation features of activities. These features are well characterized and exploited by **SensCapsNet-1** for constructing feature map. Moreover, **SensCapsNet-1** could avoid the loss of information and it can be able to handle data fusion from multiple sensors. Activities such as *Brushing, Mixing, Wiping, Sweeping floor, Turning shoulder, Turning wrist, Turning knee, Turning haunch, Cycling* are also well recognized with F-1 score ranging from 80.58% to 89.72%. These high results are obtained thank to distinctive movements. In term of F-1 score, performance of **SensCapsNet-1** is 78.73% in case of *Walking*, 75.09% with *Down-stair* and 70.37% with *Running*.

In contrast, some other activities such as *Peeling, Turning ankle* are significantly misclassified. F1-score achieved around 50% as sensors on smartwatch might possibly be noisy and interfering with the sensors of e-Shoes. Fig. 7 shows the confusion matrix obtained by **SensCapsNet-1**. We could see that the activity *Peeling* was confused with many other activities for example *Brushing*. It could be explained by the fact that while performing both activities (*Brushing* and *Peeling*), the subject does move mostly the hand but not the lower body part. The acceleration and gyroscope data from both activities are then quite similar.

Beside, *Null* activities are easily misclassified with other activities as obviously they contain significant noises which make the precision, recall, and F1-score of *null* down to around 48.45%. In fact, the recognition of activities belonging to the *Null* class is very challenging because they could include any arbitrary activity (a wide range of activities) that the subject performed that is irrelevant to the pre-defined scenario.

4) IMPACT OF USING SINGLE OR MULTIPLE SENSORS ON RECOGNITION PERFORMANCE

We have analyzed the performance of **SensCapsNet-1** on **19NonSens** dataset with the use of multiple sensors (Smartwatch and e-Shoes). We now investigate the contribution of each sensor in activity recognition. Table 7 shows the

TABLE 7. Performance of **SensCapsNet-1** when using single or multiple sensors.

Sensors	Precision	Recall	F1-score
SmartWatch	77.4 ± 1.5	78.7 ± 1.4	77.4 ± 1.5
e-Shoes	64.7 ± 1.0	63.9 ± 0.8	62.7 ± 1.0
SmartWatch + e-Shoes	$\mathbf{78.0 \pm 1.2}$	$\mathbf{78.8 \pm 0.9}$	$\mathbf{77.7 \pm 1.0}$

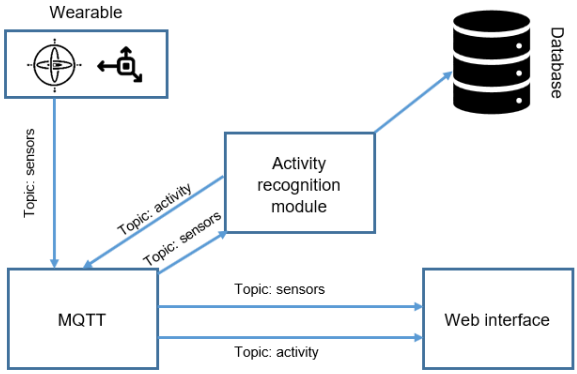


FIGURE 10. Main components of the application.

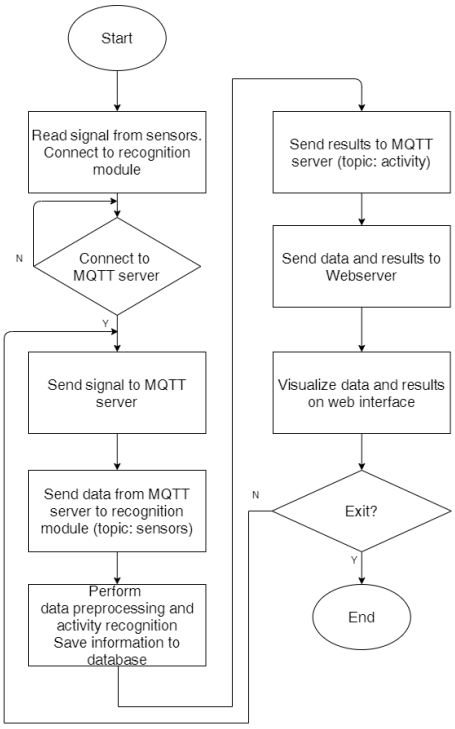


FIGURE 11. System flowchart.

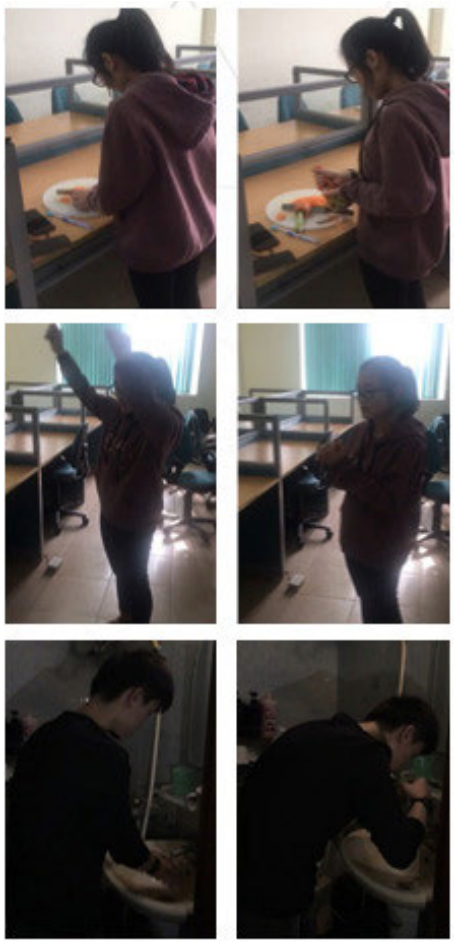
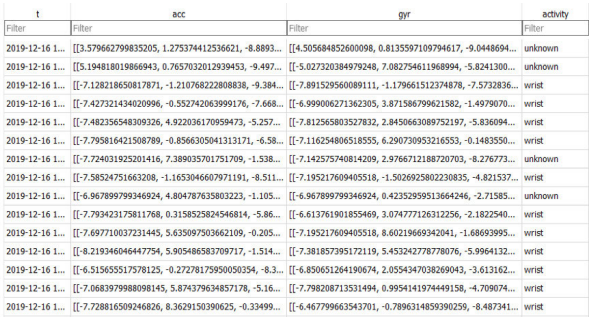


FIGURE 12. Some images of subjects performing activities to test our application.

FIGURE 13. Web interface of the application with the activity bar in the bottom. The recognized activity is marked in green.

performance of **SensCapsNet-1** achieved while using either Smart-watch or e-Shoe (the first and second row of Tab. 7) and using both sensors (the last row of Tab. 7). As can be seen, using the signal from Smart-watch could obtain the performance (77.4% F1-score), which is as high as the use of both Smart-watch and e-Shoe combination (77.7% F1-score). This could be explained by the fact that in all activities, even for activities performed mostly by lower body part, hands always involved in activity implementation and the hand movement of each activity could be quite distinctive for recognition.

The use of single sensor modality of e-Shoe, in contrast, significantly reduces the recognition accuracies (the second



	t	acc	gyr	activity
	Filter	Filter	Filter	Filter
1	2019-12-16 1...	[[3.579662799835205, 1.275374412536621, -8.8893...	[[4.505684852600098, 0.8135597109794617, -9.0448694...	unknown
2	2019-12-16 1...	[[5.194818019866943, 0.7657032012939453, -9.497...	[[-5.027320384979248, 7.082754611968994, -5.8241300...	unknown
3	2019-12-16 1...	[[-7.128218650817871, -1.210768222808838, -9.384...	[[-7.891529560089111, -1.179661512374878, -7.5732836...	wrist
4	2019-12-16 1...	[[-7.427321434020996, -0.552742063999176, -7.668...	[[-6.999006271362305, 3.871586799621582, -1.4979070...	wrist
5	2019-12-16 1...	[[-7.482356548309326, 4.922036170959473, -5.257...	[[-7.812565803527832, 2.8450663089752197, -5.836094...	wrist
6	2019-12-16 1...	[[-7.795816421508789, -0.8566305041313171, -6.58...	[[-7.116254806518555, 6.290730953216553, -0.1483550...	wrist
7	2019-12-16 1...	[[-7.724031925201416, 7.389035701751709, -1.538...	[[-7.142575740814209, 2.9766712188720703, -8.276773...	unknown
8	2019-12-16 1...	[[-7.58524751663208, -1.1653046607971191, -8.511...	[[-7.195217609405518, -1.5026925802230835, -4.821537...	wrist
9	2019-12-16 1...	[[-6.967899979346924, 4.804787635803223, -1.105...	[[-6.967899979346924, 0.4235295951366246, -2.71585...	unknown
10	2019-12-16 1...	[[-7.793423175811768, 0.3158525824546814, -5.86...	[[-6.613761901855469, 3.074777126312256, -2.1822540...	wrist
11	2019-12-16 1...	[[-7.697710037231445, 5.635097503662109, -0.205...	[[-7.195217609405518, 8.60219669342041, -1.68693995...	wrist
12	2019-12-16 1...	[[-8.219346046447754, 5.905486583709717, -1.514...	[[-7.381857395172119, 5.453242778778076, -5.9964132...	wrist
13	2019-12-16 1...	[[-6.515655517578125, -0.2727817590050354, -8.3...	[[-6.850651264190674, 2.0554347038269043, -3.613162...	wrist
14	2019-12-16 1...	[[-7.068397998809814, 5.874379634857178, -5.16...	[[-7.798208713531494, 0.9954141974449158, -4.709074...	wrist
15	2019-12-16 1...	[[-7.728816509246826, 8.362915039062, -0.33499...	[[-6.467799663543701, -0.7896314859390259, -8.487341...	wrist

FIGURE 14. Logging information stored in SQLite database.

TABLE 8. Recognition results obtained for 5 subjects (S1 to S5) when using developed application.

Subjects	Precision	Recall	F1-score
S1	0.86	0.83	0.83
S2	0.87	0.82	0.84
S3	0.87	0.88	0.86
S4	0.87	0.82	0.82
S5	0.94	0.93	0.93

row of Tab. 7). Looking at the confusion matrix output of e-Shoe (fig. 8), we can see a majority of indoor activities with less leg movement such as slicing and peeling have very low recognition rate (20-35%), while outdoor activities such as turning ankle, turning haunch, and walking are significant higher accuracies (87 to 96%) than others. However, there is an exception for going-down stairs which is low accuracy of 26% as it is very often misclassified as running, turning knee, and kicking, which leads to the accuracy average of e-Shoe (62- 65%) is significantly lower than the accuracies of multi-modality sensing of Smart-watch, or the combination of Smart-watch and e-Shoe.

VI. DEVELOPING A REAL-TIME LIFE LOGGING APPLICATION USING THE PROPOSED SensCapsNet

Based on the model trained for **19NonSens** dataset, we have built a real-time human activity and logging application. Fig. 10 shows the main modules of the application including wearable sensors, activity recognition module, database and interface. The sensor used in the application is Samsung Gear G2. In order to communicate between different modules of the application, MQTT, a machine-to-machine (M2M) connectivity protocol, is chosen [38]. A database is designed and deployed using SQLite to log the information of working sessions while a web page is built by using NodeJs framework. The flowchart of the application is illustrated in Fig. 11.

After deploying our application, 5 volunteer subjects are asked to test our application (see Fig. 12). As in the application, subjects wear only the smart watch, we ask them to perform 5 different upper body activities including brushing, peeling, turning shoulder, slicing, turning wrist in 2 minutes. Some snapshots of the web interface and database are shown in Fig 13 and Fig. 14. The recognition accuracy obtained is shown in Tab. 8. The Precision, Recall and F1-score obtained for all 5 subjects are greater than 80% for 5 upper body activities. It is worth to note that these subjects do not participate in **19NonSens** dataset building. This promising results confirm the reliability of the application. However, the number of testing subjects is still limited and the current application takes only information from only a smart watch. In the future, we aim to invite more subjects and conduct more experiments to get a full evaluation of the developed application.

VII. CONCLUSIONS AND FUTURE WORKS

In this paper, firstly, a non-obtrusive activity dataset named **19NonSens** using wearable sensor has been built. This dataset contains 19 activities collected from 12 subjects by using two devices (Samsung Gear G2 and e-Shoe). Accelerometers from smart watch and e-Shoe and gyroscope from smart watch as well as images captured by surveillance cameras have been synchronized and carefully annotated. Second, we have proposed a method for human activity recognition from wearable sensors based on capsule network **SensCapsNet**. The proposed method has been evaluated on two datasets: a subset of Opportunity and **19NonSens**. The experimental results confirms the robustness of the proposed method in comparison with two baseline deep learning-based methods. Extensive experiments have been performed in order to analyze the behavior of the proposed method for different kinds of activity as well as different information inputs. Finally, based on the proposed method, we have developed and deployed successfully a real-time human activity recognition and logging application. Different directions can be followed to improve the current work in the future. First, different dynamic routing algorithms will be investigated in order to capture the specific characteristic of the signal coming from wearable sensors. Second, the application should be deployed in embedded plate-form in order to make it usable for end-users.

REFERENCES

[1] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016786551830045X

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

[3] M. Zhang and A. A. Sawchuk, "USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors," in *Proc. ACM Conf. Ubiquitous Comput. (UbiComp)*, Pittsburgh, PA, USA, Sep. 2012, pp. 1036–1043.

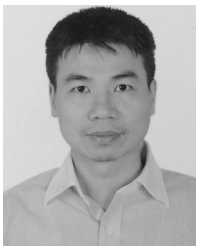
[4] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. D. R. Millan, "Collecting complex activity datasets in highly rich networked sensor environments," in *Proc. 7th Int. Conf. Netw. Sens. Syst. (INSS)*, Kassel, Germany, Jun. 2010, pp. 233–240.

[5] E. Kim, S. Helal, and D. Cook, "Human activity recognition and pattern discovery," *IEEE Pervas. Comput.*, vol. 9, no. 1, pp. 48–53, Jan. 2010.

[6] X. Liu, L. Liu, S. J. Simske, and J. Liu, "Human daily activity recognition for healthcare using wearable and visual sensing data," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Chicago, IL, USA, Oct. 2016, pp. 24–31.

[7] Y. Jia, "Diatetic and exercise therapy against diabetes mellitus," in *Proc. 2nd Int. Conf. Intell. Netw. Intell. Syst.*, Nov. 2009, pp. 693–696.

[8] J. Yin, Q. Yang, and J. J. Pan, "Sensor-based abnormal human-activity detection," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 8, pp. 1082–1090, Aug. 2008.
[9] T. Plötz, P. Moynihan, C. Pham, and P. Olivier, "Activity recognition and healthier food preparation," in *Activity Recognition in Pervasive Intelligent Environments*. Paris, France: Atlantis Press, 2011, pp. 313–329.
[10] C. Pham and P. Olivier, "Slice&Dice: Recognizing food preparation activities using embedded accelerometers," in *Proc. AmI*. Salzburg, Austria: Springer, 2009, pp. 34–43.
[11] M. Buettner, R. Prasad, M. Philipose, and D. Wetherall, "Recognizing daily activities with RFID-based sensors," in *Proc. 11th Int. Conf. Ubiquitous Comput.* New York, NY, USA: ACM, Sep. 2009, pp. 51–60.
[12] T. L. M. van Kasteren, G. Englebienne, and B. J. A. Kröse, "An activity monitoring system for elderly care using generative and discriminative models," *Pers. Ubiquitous Comput.*, vol. 14, no. 6, pp. 489–498, Sep. 2010.
[13] J. Sarkar, L. T. Vinh, Y.-K. Lee, and S. Lee, "GPARS: A general-purpose activity recognition system," *Appl. Intell.*, vol. 35, no. 2, pp. 242–259, Mar. 2010.
[14] E. M. Tapia, S. S. Intille, and K. Larson, "Activity recognition in the home using simple and ubiquitous sensors," in *Proc. Int. Conf. Pervas. Comput.* Linz, Austria: Springer, 2004, pp. 158–175.
[15] D. Avrahami, M. Patel, Y. Yamaura, and S. Kratz, "Below the surface: Unobtrusive activity recognition for work surfaces using RF-radar sensing," in *Proc. Conf. Hum. Inf. Interact. Retr. (IUI)*. New York, NY, USA: ACM, 2018, pp. 439–451.
[16] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explor. Newsl.*, vol. 12, no. 2, pp. 74–82, Mar. 2011.
[17] X. Su, H. Tong, and P. Ji, "Accelerometer-based activity recognition on smartphone," in *Proc. 23rd ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*. New York, NY, USA: ACM, 2014, pp. 2021–2023.
[18] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber, "Smartwatch-based activity recognition: A machine learning approach," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform. (BHI)*, Feb. 2016, pp. 426–429.
[19] C. Pham, N. N. Diep, and T. M. Phuong, "E-shoes: Smart shoes for unobtrusive human activity recognition," in *Proc. 9th Int. Conf. Knowl. Syst. Eng. (KSE)*, Oct. 2017, pp. 269–274.
[20] S. Bamberg, A. Y. Benbasat, D. M. Scarborough, D. E. Krebs, and J. A. Paradiso, "Gait analysis using a shoe-integrated wireless sensor system," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 4, pp. 413–423, Jul. 2008.
[21] J. Cheng, M. Sundholm, B. Zhou, M. Hirsch, and P. Lukowicz, "Smart-surface: Large scale textile pressure sensors arrays for activity recognition," *Pervas. Mobile Comput.*, vol. 30, pp. 97–112, Aug. 2016.
[22] L. T. Nguyen, M. Zeng, P. Tague, and J. Zhang, "Recognizing new activities with limited training data," in *Proc. ACM Int. Symp. Wearable Comput. (ISWC)*, 2015, pp. 67–74.
[23] 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 *Proc. IJCAI*, Buenos Aires, Argentina, Jul. 2015, pp. 3995–4001.
[24] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016.
[25] A. Murad and J.-Y. Pyun, "Deep recurrent neural networks for human activity recognition," *Sensors*, vol. 17, no. 11, p. 2556, Nov. 2017.
[26] T. Chen, Z. Wang, X. Yang, and K. Jiang, "A deep capsule neural network with stochastic delta rule for bearing fault diagnosis on raw vibration signals," *Measurement*, vol. 148, Dec. 2019, Art. no. 106857. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0263224119307146
[27] K. Suri and R. Gupta, "Continuous sign language recognition from wearable IMUs using deep capsule networks and game theory," *Comput. Electr. Eng.*, vol. 78, pp. 493–503, Sep. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0045790619301508
[28] D. Anguita, A. Ghio, L. Oneto, and X. Parra, "A public domain dataset for human activity recognition using smartphones," in *Proc. Eur. Symp. Artif. Neural Netw.*, Bruges, Belgium, 2013, pp. 24–26.
[29] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. D. R. Millán, and D. Roggen, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 2033–2042, Nov. 2013.
[30] M. Bachlin, M. Plotnik, D. Roggen, I. Maidan, J. M. Hausdorff, N. Giladi, and G. Troster, "Wearable assistant for Parkinson's disease patients with the freezing of gait symptom," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 436–446, Mar. 2010.
[31] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Tröster, "Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection," in *Wireless Sensor Networks*, R. Verdone, ed. Berlin, Germany: Springer, 2008, pp. 17–33.
[32] Y. Kim, P. Wang, Y. Zhu, and L. Mihaylova, "A capsule network for traffic speed prediction in complex road networks," in *Proc. Sensor Data Fusion, Trends, Solutions, Appl. (SDF)*, Oct. 2018, pp. 1–6.
[33] K. Duarte, Y. S Rawat, and M. Shah, "VideoCapsuleNet: A simplified network for action detection," 2018, *arXiv:1805.08162*. [Online]. Available: http://arxiv.org/abs/1805.08162
[34] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Proc. ICANN*, Espoo, Finland, Jun. 2011, pp. 44–51.
[35] S. N. F. Sabour and G. E. Hinton, "Dynamic routing between capsules," in *Proc. NIPS*, Long Beach, CA, USA, Dec. 2017, pp. 3859–3869.
[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[37] P. Lukowicz, G. Pirkl, D. Bannach, F. Wagner, A. Calatroni, K. Förster, T. Holleczek, M. Rossi, D. Roggen, G. Tröster, J. Doppler, C. Holzmann, A. Riener, A. Ferscha, and R. Chavarriaga, "Recording a complex, multi modal activity data set for context recognition," in *Proc. 23rd Int. Conf. Archit. Comput. Syst. (ARCS)*. Hannover, Germany: VDE, 2010, pp. 1–6.
[38] *MQTT*. Accessed: Mar. 30, 2020. [Online]. Available: http://mqtt.org/



CUONG PHAM received the B.S. degree from Vietnam National University, in 1998, the M.S. degree from New Mexico State University, USA, in 2005, and the Ph.D. degree from Newcastle University, U.K., in 2012, all majors in computer science. He was a Visiting Professor with the University of Palermo, Italy and a Marie Curie Research Fellow with Philips Research, Eindhoven, The Netherlands. He is currently an Associate Professor of computer science with the Posts and Telecommunications Institute of Technology (PTIT) and a Visiting Research Scientist with VinAI Research. His main research interests are ubiquitous computing, wearable computing, human activity recognition, and machine learning/deep learning.

SON NGUYEN-THAI received the B.S. degree in computer science and security from the Posts and Telecommunications Institute of Technology. He is currently a Research Engineer with the AI Tokyo Lab. His research interests are deep learning, human activity recognition, and wearable computing.

HUY TRAN-QUANG received the B.S. degree in control and automation from the Hanoi University of Science and Technology. He is currently an AI Engineer with Vietnam Post and Telecommunication-Information Technology. His research interests are deep learning and computer science.

SON TRAN received the Ph.D. degree in computer engineering from The University of Texas at El Paso, in 2000. He was a Postdoctoral Fellow with Stanford University before joining New Mexico State University (NMSU) in 2001, where he is currently a Full Professor. His main research interests are in knowledge representation and reasoning, commonsense reasoning, logic programming, automated planning, multi-agent system, and the exploitation of knowledge representation techniques in practical applications.

HAI VU received the B.E. degree in electronic and telecommunications and the M.E. in degree in information processing and communication from the Hanoi University of Science and Technology (HUST), in 1999 and 2002, respectively, the Ph.D. degree in computer science from Osaka University, Japan, in 2009. He has been a Lecturer and a Researcher with the Department of Computer Vision, MICA International Research Institute (HUST-Grenoble INP), since 2012. His current research interests are in computer vision, pattern recognition, particularly, applying these techniques in agricultural engineering, medical imaging, and human–computer interactions.

THANH-HAI TRAN received the degree in information technology from the Hanoi University of Science and Technology (HUST), in 2001, the M.S. and Ph.D. degrees in imagery vision robotic from Grenoble INP, France, in 2002 and 2006, respectively. She is currently a Lecturer/Researcher with the Department of Computer Vision, School of Electronics and Telecommunications, International Research Institute in Multimedia, Information, Communication and Application, HUST. Her main research interests are visual object recognition, video understanding, human–robot interaction, and text detection for applications in computer vision.

THI-LAN LE received the degree in information technology from the Hanoi University of Science and Technology (HUST), Vietnam, the M.S. degree in signal processing and communication from HUST, the Ph.D. degree in video retrieval from INRIA Sophia Antipolis, France, in 2009. She is currently a Lecturer/Researcher with the Department of Computer Vision, HUST. Her research interests include computer vision, content-based indexing and, retrieval, video understanding, and human–robot interaction.