Deep learning for Heterogeneous Human Activity Recognition in Complex IoT Applications

Mohamed Abdel-Basset, Hossam Hawash, Victor Chang, Ripon K. Chakrabortty and Michael Ryan

***Abstract*— With continued improvements in wireless sensing technology, the notion of the Internet of Things (IoT) has been widely adopted and has become pervasive owing to its broad applications in scenarios such as ambient assisted living, smart healthcare, and smart homes. In that regard, Human Activity Recognition (HAR) is a vital element of intelligent systems to undertake persistent surveillance of human behavior. Due to the omnipresent impact of smartphones in each person's life, smartphone inertial sensors are used as a case study for this research. Most of the conventional approaches regard HAR as a time series classification problem; yet, the accuracy of recognition degrades for heterogeneous sensors. In this paper, we investigate encoding sensory heterogeneous HAR (HHAR) data into three-channel image representation (i.e. RGB), hence treat the HHAR task as an image classification problem. Since present convolutional network models are computationally heavy when deployed in the IoT environment, we propose a lightweight model image encoded HHAR, calledmulti-scale image encoded HHAR (MS-IE-HHAR). The model employs a Hierarchical Multi-scale Extraction (HME) module followed by an Improved Spatial-wise and Channel-wise Attention (ISCA) module to form the main architecture of the model. The HME module is formed by a group of residually connected shuffle group convolutions (SG-Conv) to extract and learn image representations from different receptive fields while reducing the number of network parameters. The ISCA module combines a lightweight spatial-wise attention (SwA) block and an improved channel-wise attention (CwA) module to enable the network to pay instructive attention to spatial correlations as well as channel interdependency information. Finally, two widely available HHAR public datasets (i.e. HHAR UCI, and MHEALTH) were used to evaluate the performance of the proposed models with accuracy over 98% and 99%, respectively, demonstrating the model superiority for modeling HAR from heterogeneous data sources.**

***Index Terms*— Human Activity Recognition; Time series Imaging; Deep Learning; sensors; Internet of Things**

I. Introduction

THe rapid development in the Internet of Things (IoT) tools and smart systems can introduce several types of real-time data (e.g., cameras' images, TV videos, and speech records, WIFI data, sensor measurements). Among them, body sensor embedded in most of the smart appliances (i.e., smart-phones, smart-watches) [1] have obtained increased attention from a wide range of researchers in the field of ubiquitous computing especially human activity recognition (HAR) [2]. Recognizing human behavior from integral sensors of inexpensive smartphones has shown its effectiveness in various applications, such as smart healthcare systems, intelligent home systems, and human-computer interaction [3,4]. As a consequence of the continued increase in the number of IoT devices, a massive amount of data is generated and is potentially available for data analysis, which promotes the development of an IoT data marketplace, in which bunches of heterogeneous data coming from different IoT sources are directed instantaneously to several IoT providers/clients and are measured to offer a novel solution for identity administration and data monetization. However, developing new IoT applications such as HAR is still challenging owing to the heterogeneity of sensory data, which must be taken into account to develop new business models for HAR by leveraging IoT data monetization, and mechanization of business plans. Therefore, the concept of heterogeneous HAR (HHAR) has emerged as a challenging task for developing such a business model, and facilitate its applicability in several IoT applications such as smart cities, smart homes, and smart healthcare.

A diversity of studies proposed for HAR tasks based on different sensors, vision-dependent HAR, apply computer vision models to learn human activities from images and videos [5]. However, these methods raise privacy issues, and their performance is subject to illumination quality. On the other hand, WIFI-dependent approaches perform activity recognition through fluctuations of WIFI signals caused by human movements captured with by one or more signal descriptors (e.g., channel state information, received signal strength indicator, and angle of arrival). Yet, these approaches can only detect human activities in a limited zone [6]. Additionally, sensor-dependent methods are broadly adopted for HAR utilizing data produced from wearable sensors (e.g., data glove). However, such sensors are expensive and intrusive. Recently, constant improvements in smart portable devices that comprise numerous robust sensors (Accelerometer (AM), Gyroscope (GY), and Magnetometer (MG)) means that the time series generated from these sensors can be effectively exploited for HAR [7] because they do not involve any additional expense and are non-intrusive.

Conventionally, HAR based on smart-phones sensors is regarded as a classification for multivariate time series that can be conducted in two ways: traditional approaches or deep learning approaches. In conventional approaches, a sophisticated feature engineering approach is applied yo time series data to capture the vigorous signal representations, known as features that are subsequently fed into machine learning (ML) for classification [8]. On the other hand, supervised and unsupervised deep learning approaches are employed for both feature extraction and activity classification [9], and show outstanding performance with homogeneous data

M. Abdel-Basset and **H. Hawash** are with the Department of Computer Science, Zagazig University, Zagazig, 44519 Egypt. (mohamed.abdelbasset@fci.zu.edu.eg; hossamreda@zu.edu.eg)
V. Chang with Teesside University, Middlesbrough, UK TS1 3BA (victorchang.research@gmail.com)
R. K. Chakrabortty and M. Ryan are with the Capability Systems Centre, School of Engineering and IT, UNSW Canberra; Australia, (r.chakrabortty@adfa.edu.au; m.ryan@adfa.edu.au)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3038416, IEEE Internet of Things Journal

sources. However, these approaches suffer from performance degradation with heterogeneous data sources, which motivates this work to leverage deep learning approaches for fusion data from heterogeneous smart-phones sensors for HAR. Owing to great success achieved by deep learning in computer vision tasks, the researchers gives more attention for transforming time series data into an image representation such as Recurrence Plot (RP), Markov Transition Field (MTF), and Gramian Angular Fields (GAF) [10]. Consequently, the task of sensor-based HAR could be addressed image classification problem by encoding sensor data into efficient image representation which is extensively investigated in this study for the first time.

In this paper, three time-series imaging techniques are adapted to transform heterogeneous time series into a three-channel image representation. For this graphical representation, two advanced multi-stream deep learning architectures were introduced for modeling and learning time-series information embedded within the generated images. In the first model, we propose a novel dual-channel architecture similar to ALexNet, in which we attach the convolutional layer with a multi-head attention module. The second model comprises six dense convolutional blocks with channel squeezing in between. In this study, six different experiments were conducted. The first two experiments evaluated the proposed models on a generated RP representation. The remainder of the experiments were conducted using two versions of GAF representation. The goal of these experiments is to reveal the information content and representation power of each image encoding and model for heterogeneous HAR.

A. Challenges

Owing to the dynamic and open nature of IoT environments, the subsequent key challenges need to be addressed in studying HAR*:*

- Most of the conventional ML approaches existing studies use handcrafted features of HAR data (e.g., time and frequency statistics), which result in an unacceptable performance of composite bodily activities in real-world scenarios. Additionally, they fail to learn from high-dimensionality and enormous volume of data that often exists in IoT environments.
- Heterogeneity in data sources is often presented in the IoT environment, and it limits learning techniques from realizing optimal performance due to variability in the nature of data from different sensors.
- Limited research studies have overcome the inefficiency of physical inclusion of 2-D images for heterogeneous time-series generated from IoT devices by investigating the encoding of 3D time-series into image representation. .
- Current CNN-based approaches have limited performance on challenging images (i.e. RP, GAF, MTF), which necessitates a significant increase in the model's width or depth to gain satisfactory performance improvement. The subsequent heavy computation required severely limits their deployment in a smart or IoT environment.
- The wide range of the current CNN-based approaches for image classification rarely exploit multiscale information for time-series images and hence do not wholly exploit the hierarchical representations.

An efficient and effective approach is required that can practically address and solve the above challenges by using deep learning is to lessen the number of computational transactions and the parameters of the network. In view of this, we propose a multi-scale image-encoded HHAR (MS-IE-HHAR) to solve these challenges by integrating multiscale learning, residual learning, and novel attention techniques in a single unified framework.

B. Contributions

This study presents a novel and lightweight framework that effectively processes and classifies human activities generated by heterogeneous sensors embedded in various IoT devices. The primary contributions of this work are:

- We redesign time-series encoding techniques that proficiently convert raw inertial time-series values to pixel values, and encode the 3D time-series of human activities into three-channel images that overcome the heterogeneity in sensory data by representing it as an activity image. Thus, the generated image could be subsequently be passed to CNN-based architectures to extract features and classify them.
- A novel hierarchical multi-scale feature extraction (HME) module is used to capture discriminative time-series encoded features at several spatial scales and positions. Importantly, the multi-scales in our model denote various accessible receptive fields instead of multi-scale images. The design of HME employed shuffled group convolution (SG-Conv) to reduce network parameters and lessen the computational burden. Consequently, the proposed HME yields a lighter and more robust model.
- A lightweight improved Spatial-wise and Channel-wise Attention (ISCA) module integrates novel spatial-wise attention (SwA) and novel channel-wise attention (CwA) to assist the network to focus on valuable information during learning.
    - ➢ The SwA decreases the cross-channel and spatial interactions via computationally efficient depthwise separable convolutions (DW-Conv).
    - ➢ The empirical evaluation demonstrates the efficiency of the SwA and CwA blocks are more efficient.

C. Paper Organization

The remaining of the paper is structured as follows: Section 2 introduces a literature review and associated works on heterogeneous activity recognition and time series imaging. The proposed frameworks and principles incorporated are described in detail in Section 3. Section 4 details the specification of the experiments, datasets, evaluation measures, results, and analysis. Section 5 introduces the limitation of this study and proposes some possible future works. Finally, the conclusion of the study and the intended future direction is discussed in Section 6.

II. RELATED WORKS

The transformation of time series data into image representation is a useful technique that plays a significant role in various multi-modality classification problems. For instance, the authors in [11] studied the classification of time series data by converting time series into RP representation and then applied a graphical descriptor (i.e., Gabor, Local Binary Patterns) to capture encoded texture patterns that were subsequently classified with Support Vector Machine (SVM). Lu et al. [12] investigated activity recognition by transforming AM data into an RP image representation and then utilized a residual network (ResNet) to learn activity information from generated images. Wang et al. [13] exploited a tiled Convolutional Neural Network (CNN) for the classification of GAF and MTF images encoded from time-series signals as images. Estebsari et al. [14] proposed a framework for load forecasting based on imaged time series in the form of RP, MTF, and GAF—they explored the effectiveness of each imaging technique on CNN classification performance. Setiawan et al. [15] investigated the HAR task by adopting a multi-channel CNN pre-trained on the AlexNet architecture for modeling time series temporal information embedded into a 3D image using GAF. Wu et al. [16] encoded AM and GY data into GAF images, which were later passed to a conventional CNN for learning and classification of activity information. Additionally, images can be produced by conducting various image processing (i.e., SURF, BRISK), or computing the distance between temporal data points. For instance, Silva et al. [17] introduced a method for classifying sequential data by expanding RP by utilizing compression distance, which enables learning of temporal patterns encoded in RP images for recognition or classification tasks.

Qin et al. [30] proposed a modified GAF encoding for heterogeneous HAR, where the global GAF was adapted to discriminate between various static activities since the traditional local GAF was unable to maintain structures corresponding to time series constancy. Thus, they integrated both global and local GAF in conjunction to be fused using an efficient ResNet architecture to forecast activity labels of each image. The authors also compared their proposed GAF images with images generated with conventional distance techniques (i.e., Minkowski distance, and the Chebyshev distance), but they did not indicate the reason for selecting the GAF technique for image encoding, and their model has low generalization performance on data not involved in training. On the other hand, Lu et al. [12], adopted the ResNet architecture for HAR by encoding acceleration data into standard RP or modified RP. Stisen et al. [28] adopted a random forest (RF) classifier to select frequency domain and time-domain features.

Similarly, Hammerla et al. [31] utilized the set of features but employed SVM as a classifier. However, the performance of these approaches depend heavily on the complicated feature engineering approach adopted. Bhattacharya et al. [32] proposed a stacked restricted Boltzmann machine (RBM) for fusing and learning hidden representation domain frequency features. In [33], each sensor data input was fed into a separate RBM architecture for concurrent processing; after which, the fused outputs were concatenated and passed through another stacked RBM. However, these approaches still suffered from contrastive divergence issues and were inefficient for either log-likelihood computation or loss tracking.

A full convolutional network (FCN) that comprises three blocks of one-dimensional convolution blocks is proposed in [34]. Each of these blocks contains a one-dimensional convolution layer together with the $ReLU$ activation function then tailed with batch normalization followed by a $SoftMax$ for final decision calculation. Ehatisham-ul-haq et al. [35] introduced two frameworks for heterogeneous HAR at different granularity levels using various classifiers, namely RF, Decision Tree (DT), and Neural Networks (NN). In coarse-grained approaches (CGA), the activity recognition was conducted without including any relevant activity information. On the other hand, in fine-grained approaches (FGA), all activity examples with similar contextual information are combined and given the same class label. This work, however, did not consider the relationship between various humans in terms of dynamic and static activities.

There have also been numerous studies regarding multi-modal source fusion and heterogeneous data processing. For instance, for action recognition, Kong et al. [20] proposed a novel framework for learning and extraction of the heterogeneous pattern from RGB-D or video data by fusing both private and shared space separately. In [21], Wang et al. organized dynamic images generated from the RGB-D videos passed to CNN to capture the interactive multi-modal features. Simonyan et al. [22] introduced a two-phase approach using dual-channel CNN architecture: in the first phase, each channel learns and extract spatial and temporal features of video streams after encoded into RGB representation; in the second phase, the calculated SoftMax scores aggregated for the final decision. However, the main drawback of such approaches is the intrusive nature of vision-based HAR .

To date most current HAR approaches, either traditional ML or deep learning models, have been adopted for homogeneous data streams and multi-modality heterogeneous fusion is a challenging area that requires more attention. Consequently, this study adopts a modified time-series imaging technique, namely RP, MTF, and GAF, for multi-modal smartphone and smartwatch sensors. We propose two novel deep network structures to learn spatial-temporal features from encoded images for HAR and to investigate the effectiveness of each imaging technique for activity recognition.

III. METHODOLOGY

This section introduces a detailed explanation of how various image encoding techniques are adopted in this work for smartphone inertial sensor data and discusses the details of the proposed deep learning approaches for activity recognition. The proposed model consists primarily of two stages: the first stage conducts time-series imaging and engineering; the second stage employs various deep learning models for feature learning and activity classification.

A. Heterogeneous Time-series Imaging for HAR

Unlike some preceding studies [43], in which time-domain signals are purely planned on a graph and subsequently encoded

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3038416, IEEE Internet of Things Journal

into a graphical format to be processed using a CNN architecture. This study chiefly contributes to the encoding value of time-series data into pixel form to perform HAR based on image representation of time-series. Unlike conventional techniques that assemble the time-series from a tri-axial sensor into a two-dimensional (2-D) matrix that has a longitudinal size of $3 \times N$ (i.e. N is the number of records), the proposed encoding mechanism is much improved by projecting the sensory values into a three-dimensional (3-D) image.

1) Recurrence Plot

RP is [12] an illustrative tool utilizing recurrence to analyze non-linearity of data points within a phase space sophisticated states of the dynamic system. RP encompasses distinctive limited-scale features (i.e., dots and lines), while wide-scale patterns can be visually represented by regular and periodic shapes. The recurrence matrix (RM) for specific trajectory data instances $Ts = \{Ts_1, \ldots. Ts_N\}$ can be calculated as shown in equation (1).

$$RM_{ij} = \mu\left(\epsilon - \left\|Ts_i - Ts_j\right\|\right), \qquad i, j = 1,2, \cdots, N \qquad (1)$$

where ϵ denotes a threshold value for states count, μ represents a unit step function and $||\,.\,||$ denotes $L2$ norm function to maintain the shape of the original data curve. We employed RP to transform 3D smart-phones signals into three-channel RGB images to exploit correlation. For example, assume a segment of AM data $D_{AM} = \{x_{AM}, y_{AM}, z_{AM}\}$, where $D \in \mathbb{R}^{3\times N}$. Each row contains x, y, and z measurements for AM, GY, and/or MG data in three coordinate dimensions, respectively. In terms of mathematical formulation, samples are provided of inertial sensors data in a specific dimension (i.e. x, y, or z); $x \in \mathbb{R}^{1\times N}$ while x^j represents j-th element of x. phase-space states denoted as $st^j = (x^j, x^{j+1})$ while $st^j \in \mathbb{R}^2$. So, for HAR, the $RM \in \mathbb{R}^{(N\times 1)\times(N\times 1)}$ from sensor data is normalized state change through L_2 norm, as expressed in equation (2).

$$RM_{\boldsymbol{i},\boldsymbol{j}} = \| st^i - st^j \| \qquad (2)$$

where m and n represent the diagonal symmetricity of the computed RM around the main diagonal with zero values, which leads to a misperception of signal inclination [12] that is known as the tendency confusion problem. For instance, consider two segments of time series Ts_1 and Ts_2, where $Ts_1 = \{t_1,\ t_2, t_3\}$ and $Ts_2 = \{t_3,\ t_2, t_1\}$, assuming that Ts_1 is of increasing order and Ts_2 is of decreasing order. Other types of interrelations that characterize small-scale oscillations are not important for classifying time series. The RM of Ts_1 and Ts_2 are calculated as shown in equations (3) and (4):

$$RM(TS_1) = \begin{vmatrix} \|(t_1,t_2) - (t_1,t_2)\| & \|(t_1,t_2) - (t_2,t_3)\| \\ \|(t_2,t_3) - (t_1,t_2)\| & \|(t_1,t_2) - (t_1,t_2)\| \end{vmatrix} \qquad (3)$$

$$RM(TS_2) = \begin{vmatrix} \|(t_3,t_2) - (t_3,t_2)\| & \|(t_3,t_2) - (t_2,t_1)\| \\ \|(t_2,t_1) - (t_3,t_2)\| & \|(t_3,t_2) - (t_3,t_2)\| \end{vmatrix} \qquad (4)$$

Despite the difference in the tendency of both, the corresponding RMs are the same. Therefore, it is difficult if not impossible for the deep learning models to differentiate human activities whose time series tendencies play a key role in detecting such activities as standing, sitting, climbing and descending stairs up.

To solve the abovementioned problem, we propose an adapted version of RP (ARP) that is able to capture and present gradient direction information without any confusion. Thus, we propose to determine the tendency using the following procedures. First, a sequence is transformed into phase space, then the subtraction and $L2 - norm$ procedures are carried out to quantify the vectors of state difference and the values of pixels of RP image correspondingly. Second, every vector of state difference is summed discretely, and the polarity of the resultant values are captured to establish a polarity mask with identical dimensions as an RP image. Third, the polarity mask is multiplied by the RP image to obtain the polarity-aware RP image. The final computation of the RM is formulated in equation (5).

$$RM_{\boldsymbol{i},\boldsymbol{j}} = \frac{\sum(st^i - st^j)}{|\sum(st^i - st^j)|}\left\|st^i - st^j\right\| \qquad (5)$$

where $\|.\|$ represents the L2-norm, and $|.|$ denotes the function for computing the absolute values.

We can then encode 3D heterogeneous sensors signals into three RP matrixes, which can be combined to form a new 3D matrix M that is subsequently normalized in accordance with the function formulated expressed in equation (6).

$$IMG = \frac{RM - min(RM)}{max(RM) - min(RM)} \qquad (6)$$

where IMG denotes the normalized 3D matrix that is used subsequently for image generation. The calculation of a minimum of M is performed ignoring zero values.

2) Markov Transition Field

MTF [13] was proposed to calculate the changeover statistics and translates these statistics into an image representation. All time-series values are divided into a static number of states by detecting quantile boxes and by presuming that every box represents a certain state. For example, given the time-series $Ts = \{Ts_1, \ldots. Ts_n\}$ consisting of n values, Ts is divided into Q quantile boxes to form $\widehat{Ts} = \{\grave{T}s_1, \ldots. \grave{T}s_n\}$ such that $\grave{T}s_i$ represents the number of the boxes of Ts_i. Thus, even though $\grave{T}s$ contains n elements, it only contains Q different values $\{q_1, q_2, \ldots, q_Q\}$. Then, the count of transitions between every single state is computed and normalized to be employed as a feature to describe the behavior of the time series. An adjacency matrix M with size $Q \times Q$ is computed from the counted transitions among quantile boxes. Thus, every element $M_{i,j}$ is assigned a value to represent the amount that q_i is tailed with q_j in Ts. A final normalization operation is performed to make $\sum_j M_{i,j} = 1$. however, based on experimental observation, we notice that applying such normalization causes great information loss. Specifically, the temporal interrelationships are lost due to a lack of capturing time-series information relevant to time periods once the transitions occurred.

Consequently, to tackle this problem, we propose to employ the normalized counting to construct a time-aware MTF matrix with a size of $N \times N$, wherein N represents the number of time-series elements. The MTF matrix is established by assigning every element at position i, j the count of state changes from time-period i to the time period j. The matrix calculation is conducted using equation (7).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3038416, IEEE Internet of Things Journal

$$MTF_{i,j} = M_{T\grave{s}_i, T\grave{s}_j} \tag{7}$$

The thought-provoking portion of MTF matrices is that the transition statistics in several phases are encoded into an image representation. For instance, $M_{i,j}$ where $|i - j| = 1$ characterizes the transition procedure over the time axis which exhibits a single unit of change. Further, increasing the number of units to 2 i.e. $|i - j| = 2$ enables the detection of the transition procedure within two units of time. Thus, the primary diagonal signifies the possibility of remaining in a specific state. Such information is important and not easy to acquire from the out-of-sensor time series, which in turn provides evidence that employing MTF representation individually or in conjunction with other representations might enable developing HAR systems in IoT environments. This is experimentally validated in later sections.

3) *Gramian Angular Fields for HAR data*

Gramian Angular Fields (GAF) [13] is a technique for transforming time-series data from traditional Cartesian coordinate systems into polar coordinate representation by encoding the normalized data values as the angular cosine. Assuming a time series $X = \{x1, x2, \ldots\ldots xn\}$ with length n of samples, the normalization function for time series into the interval [-1,1] can be computed as shown in equation (8).

$$\tilde{x}_{-1}^{i} = \frac{(x_i - max(X)) + (x_i - min(X))}{max(X) - min(X)} \tag{8}$$

Then, the angular cosine function is applied to the quantified time series $\tilde{\mathrm{x}}$ in order to encode them into polar coordinates where the radius denotes time steps of data. This operation can be computed as formulated in equation (9).

$$\begin{cases} \phi = arccos(\tilde{x}_i), -1 \leq \tilde{x}_i \leq 1, \ \tilde{x}_i \in \tilde{\mathrm{x}} \\ r = \frac{t_i}{n} \ t_i \in N \end{cases} \tag{9}$$

Then, the preserved temporal relations encoded in polar coordinate representation can be exploited by calculating the sum or difference between scaled time series points at different time intervals. Hence, the calculation of Gramian Angular Summation Field (GASF) and Gramian Angular Difference Field (GADF) is calculated with equation (10) and equation (11), respectively.

$$GASF(i,j) = cos(\phi_i + \phi_j) = \tilde{\mathrm{x}}'.\tilde{\mathrm{x}} - \sqrt{I - \tilde{\mathrm{x}}^2}'.\sqrt{I - \tilde{\mathrm{x}}^2} \tag{10}$$

$$GADF(i,j) = sin(\phi_i - \phi_j) = \tilde{\mathrm{x}}'.\sqrt{I - \tilde{\mathrm{x}}^2} - \tilde{\mathrm{x}}.\sqrt{I - \tilde{\mathrm{x}}^2}' \tag{11}$$

In the adapted GAF images (i.e. AGADF or AGASF) for HAR data, we normalize dynamic activity measurements with equation (8), and to overcome static activity stability in polar coordinate, we normalize static activity measurements with equation (12).

$$\tilde{x}_{-1}^{i} = \frac{(x_i - Q3(X)) + (x_i - Q1(X))}{Q\,3(X) - Q1(X)} \tag{12}$$

where $Q\,3$ and $Q1$ denote the third and the first quartiles, respectively. For convenience, we refer to the GAF image generated using equation (8) as dynamic GASF (D-GASF) or dynamic GADF (i.e. D-GADF). In the same manner, we refer to the GAF image generated using equation (12) as static GASF (S-GASF) or static GADF (i.e. S-GADF).

4) Mixed Representation

To take the advantage of multiple-image encoding techniques, we propose to concatenate the extracted features from GADF and GADF images, and subsequently pass the concatenated features for the classifier. We refer to this mixed image representation as Mix1. Similarly, we create Mix2 representation by concatenating the feature representation acquired from GADF, GASF, and MTF images.

B. *Proposed Deep Learning Model*

A variety of deep learning studies have demonstrated the efficiency of CNN-based models for time-series image classification [23-27]. Nevertheless, the standard convolution architecture can only extract features through a static receptive field that limits shallow layers from acquiring and learning a comprehensive set of spatial representations which, in turn, reduces the network capability in discriminating various activity representation in different kinds of the before-mentioned images. To tackle this challenge, we propose the MS-IE-HHAR model presented in Fig. 1, in which a hierarchical multiscale extraction (HME) module is introduced to fine-tune the process of extracting spatial representations. Additionally, an improved Spatial-wise and Channel-wise Attention (ISCA) Module is proposed to selectively improve the features by encoding the spatial importance and the channel importance into learnable parameters.

1) *Hierarchical Multiscale Extraction module*

Multiscale information is indispensable for numerous computer vision applications. The efficient extraction of this information has a vital role in developing a robust and accurate deep learning model. Further, we place emphasis on solving the efficiency challenge that is principally required in real-world IoT applications. For balancing the computational complexity and the performance, several studies [24], [26], [47] sought to

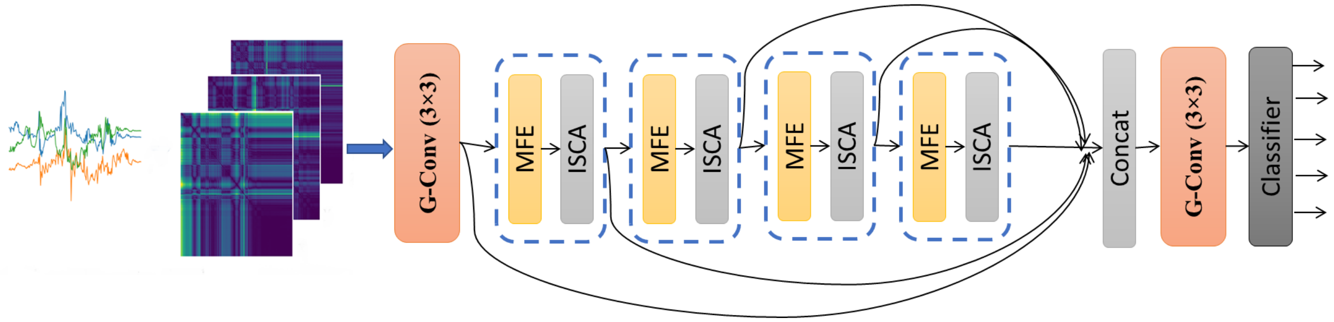

Fig. 1. The Architecture of the proposed MS-IE-HHAR consisting of group convolution(G-Conv) at first, followed by four layers, each layer contains the MFE module followed by ISCA module. The output of these layers is directed to a concatenation layer, after which the final computation performed with G-Conv (3×3). The final decision is calculated at classifier layer.

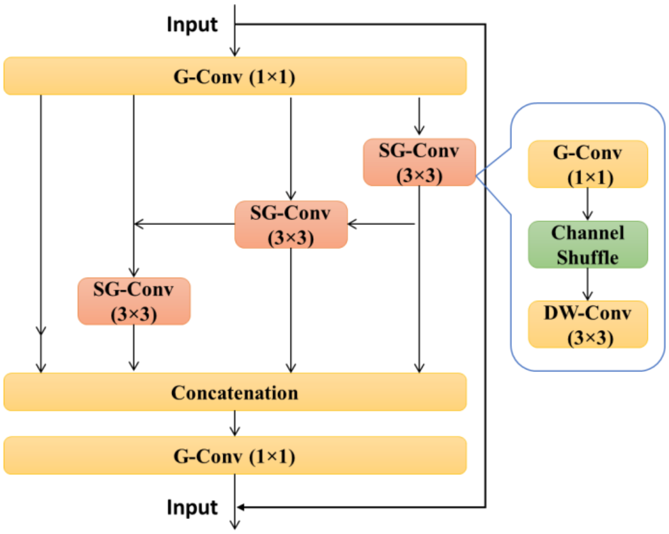

Fig. 2. The architecture of the proposed hierarchical multiscale extraction (HME) module

develop multiscale modules. Nevertheless, they suffer from limited receptive fields, and also are computationally complex.

The architecture of the proposed HME module is presented in Fig. 2, and consists of several type layers. First, the input is fed into 1 × 1 G-Conv to reduce its dimensionality and to decrease the processing workload and subsequently pass the output to the next four parallel paths. The leftmost most path is designated using the average pooling (AVP) layer to enable extraction and reuse of important features through training. Other paths are established using shuffled group conv (SG-Conv) [48] in which a channel shuffling procedure takes the feature map incoming from the preceding G-Conv layer and split the channels in every group into numerous subgroups, and pass each group to the following layer as diverse subgroups. This enables the construction of a robust architecture with numerous G-Conv layers.

To assure that the proposed propose MS-IE-HHAR is lighter, the G-Conv layers were employed to construct the HME module (see Fig. 2). This alleviates the load incurred by a complex architecture established by stacking several blocks. The G-conv with 1 × 1 filters acquire feature maps f_i from the received input, and process them at different scales using SG-Convs. Then, the production of a certain path is passed to the concatenation layer and simultaneously used as an input to the succeeding path via an element-wise addition. This process is duplicated multiple times until the feature maps generated from all paths are extracted. This operation is formulated as in equation (13).

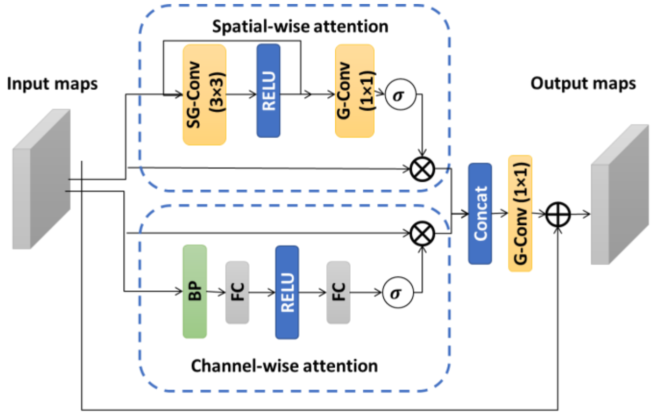

Fig. 3. The architecture of the proposed an Iimproved Spatial-wise and Channel-wise Attention (ISCA) Module

$$F_i = \begin{cases} AVP(f_i) & i = 1 \\ SG - Conv(f_i) & i = 2 \\ SG - Conv(f_i + F_{i-1}), & 2 < i \leq 4 \end{cases} \qquad (13)$$

where $AVP(f_i)$ represents the procedure of the leftmost path, and F_i characterizes the calculated output map.

Once the feature extraction is complete in all paths, the feature maps of different scales are fused using concatenation operation which does not cause any information loss compared with other fusion techniques [35]. Then, the concatenated maps are fed into the 1 × 1 Conv to enable the effective processing of features. In order to maintain the intrinsic image representations, the outcome of the HME module is subsequently combined by the original input maps via the residual connection. The HME module is therefore beneficial for exploiting visual representation at various scale-spaces.

2) *Improved Spatial-wise and Channel-wise Attention (ISCA) Module*

Since the learned feature representation acquired by the HME module comprises diverse categories of information from different spatial regions and various channels, so it is critical to develop a technique to enhance the ability of the model to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3038416, IEEE Internet of Things Journal

access such valuable details. Thus, to enable the MS-IE-HHAR to focus on the appropriate features to improve its representative power, spatial correlations and channel interdependencies are taken into account to construct the ISCA module as presented in Fig. 3, in which a lightweight spatial-wise attention (SwA) and channel-wise attention (CwA) are combined.

SwA block. The main purpose of SwA block is to give emphasis to various surface-level areas and allocate attention parameters corresponding to pixels in every convolutional map. It is obvious that the time-series information encoded in the image and embraced in convolutional maps differs across various surface-level locations. The detailed information about regions such as diagonals or composite surfaces is often of great importance. On the other hand, smooth zones contain limited information. In order to develop spatial pixel-level attention, a spatial attention mechanism [26] was introduced to calculate attention maps using max and mean pooling layers to accumulate channel-wise representations. Nevertheless, these pooling layers can miss imperative channel-wise statistics and hence are inefficient. To tackle this issue, we primarily concentrate on learning the more useful spatial information by lessening the spatial-wise and channel-wise associations.

Motivated by [17], in which DW-Conv is first used to minimize such kinds of association, this work proposes the a novel spatial attention module, SwA which comprises three main part. First, PG-Conv is employed to capture channel-wise interdependence and reducing its dimension. Second, a channel shuffling procedure is applied and followed by the DW-Conv layer to reduce the cross-channel correlations and spatial correlations. Third, a subsequent PG-Conv layer was employed to produce a spatial attention map. The computational procedure of SwA is shown in Fig. 3. Given the input map, $U = \{u_1, u_2, \cdots u_k\}$, which consists of k channels, with dimensions of $w \times h$. The SwA uses the three layers to calculate a SwA mask as shown in equation (14).

$$\beta = \sigma(W_{swA}^2 \delta(W_{swA}^1 U)) \tag{14}$$

where $\sigma(\cdot)$ and the $\delta(\cdot)$denote the sigmoid and ReLU function respectively. $\beta \in \mathbb{R}^{1\times w\times h}$ represents the attained mask of SwA. Lastly, the factor β used to modulate the outcome of the SwA block as formulated in equation (15).

$$U_{SwA} = \beta \odot U \tag{15}$$

wherein the operation of element-wise multiplication is represented by " $\odot$ " and $U_{SwA} = \{\tilde{u}_1, \tilde{u}_2, \cdots \tilde{u}_k\}$, $U_{SwA} \in \mathbb{R}^{h\times w\times k}$ denotes the total outcome of the SwA block.

CwA block. The main target of CwA is to extract the second-order statistics from every channel of the incoming map U and to determine the interdependency between second-order global and local features in an integrated unit, whilst maintaining the spatial construction of the input U . This is realized by recalibrating the attained features of the HME module, which avoids treating all the channels similarly and ignoring the channel interdependence. Inspired by the fact that bilinear pooling (BP) has been shown to be an efficient tool for fine-grained visual recognition, BP is introduced at the beginning of CwA to calculate the pairwise feature interrelations. For every cell of feature map, the external product is calculated with equation (16).

$$X_i = u_c u_c^T \tag{16}$$

The BP procedure is conducted to capture the pairwise interdependence between the convolutional channels, that can be computed using the Gram matrix $G \in \mathbb{R}^{\boldsymbol{k\times k}}$**,** as in equation (17)**.**

$$G = \frac{1}{w\times h}\sum_{i=1}^{hw} u_i u_i^T = \frac{1}{w\times h} XX^T \tag{17}$$

where G represents the entire feature space representation, capturing the activation related second-order statistics. Subsequently this representation is compressed, resulting in the channel-wise information $z \in \mathbb{R}^{K\times1\times1}$. The GAP is conducted along the spatial dimension. The c-th component of z is calculated with equation (18).

$$z_c = \frac{1}{w\times h}\sum_{i=1}^{h}\sum_{j=1}^{w} G_c(i,j) \tag{18}$$

Further, channel-wise parameters (i.e. weights) are assigned to each feature map using a simple gating technique, wherever the sigmoid operation is exploited to obviously learn the channel-wise correlations as depicted in equation (19).

$$S = \sigma(W_{CwA}^2 \delta(W_{CwA}^1 z)) \tag{19}$$

where $\sigma(\cdot)$ and the $\delta(\cdot)$ have the same definition in equation (14). $W_{\mathrm{CwA}}^1 \in \mathbb{R}^{\frac{k}{r}\times k\times1\times1}$, $W_{\mathrm{CwA}}^1 \in \mathbb{R}^{k\times\frac{k}{r}\times1\times1}$ represent the parameters corresponding to each layer. The first fill connected (FC) layer preceding the $ReLU$ function uses a discount ratio r to lessen the channel of z. The succeeding layer is accountable for restoring it to its original form, then the attention parameters are calculated by the $sigmoid$ function. To that end, the production of the CwA block is adapted with S according to equation (20).

$$U_{SwA} = S \odot U \tag{20}$$

where $\odot$ is defined as in equation (15), and $U_{CwA} = \{\tilde{u}_1, \tilde{u}_2, \cdots \tilde{u}_k\}$, $U_{CwA} \in \mathbb{R}^{h\times w\times k}$ denotes the total outcome of the CwA block. Unambiguously, for every element, we have $\tilde{u}_1 = u_c \cdot s_c$.

3) Classifier layer

At the end of the MS-IE-HHAR, the concatenated output is convolved with G-Conv (3×3) and passed to the classifier layer to determine the activity class. The FC layer is adapted to convert the incoming feature maps F_m into a fine-grained representation that is suitable for forecasting the activity classes. The SoftMax function is utilized to calculate the probability score for each activity class, while the class that gains a higher probability score is considered to be the correct activity class as formulated in equations (21) and (22).

Table I. HHAR UCI dataset summary

	Accelerometer		Gyroscope	
Human Activity	**Samples**	**Percentage**	**Samples**	**Percentage**
Standing (A1)	2302681	16.10%	2454429	16.59%
Sitting (A2)	2415914	16.89%	2638035	17.83%
Biking (A3)	2481087	17.35%	2434402	16.45%
Walking (A4)	2742162	19.18%	2838738	19.18%
Stairs Up (A5)	2255764	15.77%	2330329	15.75%
Stairs down (A6)	2102272	14.70%	2102074	14.21%
Total	14299880	100%	14798007	100%

$$p = SoftMax(F_m) = \frac{exp(F_m)}{\sum_{1}^{c} exp(F_m)} \quad (21)$$

$$\tilde{y} = argmax(p) \quad (22)$$

Sooner or later, to training or evaluate our model, we need to calculate and minimize our loss function. In this paper, for better generalization, we adopt a recently introduced pairwise Gaussian loss (PLG) [24] in conjunction with Entropy loss, where the loss function computed as with equations (23) and (27):

$$g(d_{ij}) = e^{-\beta\, d_{ij}^2} \quad (23)$$

$$P(y_{ij}, d_{ij}) = \begin{cases} g(d_{ij}), & y_{ij} = 1 \\ 1 - g(d_{ij}), & y_{ij} = 0 \end{cases} = [g(d_{ij})]^{y_{ij}} - [1 - g(d_{ij})]^{(1-y_{ij})} \quad (24)$$

$$L_{PGL} = \frac{4}{N^2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} - log\, P(y_{ij}, d_{ij}) \quad (25)$$

$$L_{Entopy} = \sum(y_i\, log(\tilde{y}_i) + (1 - y_i)\ log(1 - \tilde{y}_i)) \quad (26)$$

$$L_{total} = L_{PGL} + L_{Entopy} \quad (27)$$

where $d_{ij} = \|f_i - f_j\|_2$ represents the Euclidian distance between two features f_i andf_j , and β denotes scaling constant (we set β =.05), where the two features belong to the same class indicated with $y_{ij} = 1$. Otherwise, $y_{ij} = 0$. y_i is the truthful activity label while $\tilde{y_i}$ is the model predicted class. In equation (27), we combine both loss functions in equation (25) and equation (26) as the final loss of our model. The model architecture with the highest performance has been chosen. A 0.4 Dropout layer is inserted before the concatenation layer. Furthermore, we conducted a grid-search for various hyper-parameters and identified the highest performance with: training epochs within range (60-80), batch size with a value of (64, 128), and with learning rate between (0.0001 – 0.00001).

IV. Experiments and Discussions

In this section, we discuss the relevant experimental conditions corresponding to our proposed approach, the data sets used for training and evaluation, the evaluation metrics adopted and then introduce the results of our proposed method on different datasets. As well as the experiments of our proposed model, we also compare our results with recent state-of-the-art approaches.

A. Datasets

A.1.UCI HHAR Dataset [28]: a public heterogeneity HAR dataset comprises both three-dimensional acceleration and gyroscope data aggregated from 9 subjects with four smartphones and two smartwatches. The dataset is collected in different usage situations and with a diversity of device models for achieving sensing heterogeneity anticipated in real productions. This dataset encompasses 14299880 labeled instances corresponding to six activities, four dynamic movement activities which are Stairs up, Stairs down, Walking, Biking, and two static activities which are standing and sitting as shown in Table I. The smartphone data has 200 Hz, 150 Hz, 100 Hz, and 50 Hz sampling frequency correspondingly, while each smartwatch data has 200 Hz, 100 Hz as a sampling frequency. In our experiments, we segmented data into three-second segments for all devices; then, we divide the segmented data into training and test data with 451162 and 112790 instances, respectively.

Table II. MHEALTH Dataset summary

Human Activity	Samples	percentage
Standing still (A1)	30720	8.95%
Sitting and relaxing (A2)	30720	8.95%
Lying down (A3)	30720	8.95%
Walking (A4)	30720	8.95%
Climbing stairs (A5)	30720	8.95%
Waist bends forward (A6)	28315	8.25%
Frontal elevation of arms (A7)	29441	8.58%
Knees bending (A8)	29337	8.55%
Cycling (A9)	30720	8.95%
Jogging (A10)	30720	8.95%
Running (A11)	30720	8.95%
Jump (A12)	10342	3.01%
Total	343195	100%

A.2.UCI MEHEALTH Dataset [29]: a public mobile health dataset for HAR. which comprises both three-dimensional data from multi-modality sensors namely accelerometer, gyroscope, and magnetometer data aggregated from 10 subjects with three devices attached to the chest, right wrist and left ankle of the user for measuring heterogeneity signals of various body movements expected to be adopted in real life. This dataset encompasses 343195 of labeled instances corresponding to twelve activities, as shown in Table II. The smartphone data has 50 Hz sampling frequency for all devices. In our experiments, we segmented data into five-second segments for all devices and got a total of 1318 instances. We train the model on 12 classes without merging any classes [30].

B. Evaluation Measures

HAR that depends on smartphone data can be regarded as a multi-classification problem where each activity class can be considered to be an independent set of instances. Accordingly, in this paper, we use accuracy, precision, recall, and F1-measure as evaluation metrics, which are reliable with those used in previous studies. The computation of these metrics can be calculated as presented in equations (28-31).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (28)$$

$$Precision = \frac{TP}{TP + FP} \quad (29)$$

$$Recall = \frac{TP}{TP + FN} \quad (30)$$

$$F1 - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (31)$$

Table III. Confusion matrix of proposed *MS-IB-HHAR using* ARP on HHAR **UCI** test set

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**R (%)**
A1	6325	8	15	12	21	11	98.95
A2	14	6699	16	15	14	10	98.98
A3	10	19	5930	14	26	17	98.57
A4	11	15	21	6843	133	121	95.79
A5	17	16	6	110	5726	140	95.20
A6	11	17	11	98	115	5010	95.21
P (%)	99.01	98.89	98.85	96.49	94.88	94.37	
F (%)	98.98	98.94	98.71	96.14	95.04	94.79	

Table IV. Confusion matrix of proposed *MS-IB-HHAR* using AGASF on HHAR **UCI** test set

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**R (%)**
A1	6355	6	10	3	11	7	99.42
A2	15	6717	12	3	13	8	99.25
A3	12	10	5956	14	19	5	99.00
A4	7	9	7	7005	55	61	98.05
A5	9	10	4	88	5828	76	96.89
A6	11	9	8	122	105	5007	95.15
P (%)	99.16	99.35	99.32	96.82	96.63	96.96	99.42
F (%)	99.29	99.30	99.16	97.43	96.76	96.05	99.25

Table V. Confusion matrix of proposed MS-IB-HHAR *using AGADF on HHAR **UCI** test set*

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**R (%)**
A1	6349	11	10	5	11	6	99.33
A2	15	6721	9	4	12	7	99.31
A3	19	14	5954	11	10	8	98.97
A4	5	8	3	6932	102	94	97.03
A5	2	8	5	96	5817	87	96.71
A6	3	6	1	102	111	5039	95.76
P (%)	99.31	99.31	99.53	96.95	95.94	96.15	
F (%)	99.32	99.31	99.25	96.99	96.32	95.95	

Table VI. Confusion matrix of proposed *MS-IB-HHAR* using MTF on HHAR **UCI** test set

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**R (%)**
A1	6261	31	29	19	21	31	97.95
A2	35	6628	34	28	25	18	97.93
A3	41	42	5854	27	28	24	97.31
A4	23	26	27	6772	138	158	94.79
A5	32	25	24	102	5666	166	94.20
A6	29	18	29	77	189	4920	93.50
P (%)	97.51	97.90	97.62	96.40	93.39	92.53	
F (%)	97.73	97.92	97.46	95.59	93.79	93.01	

Table VII. Confusion matrix of proposed *MS-IB-HHAR* using Mix1 on HHAR **UCI** test set

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**R (%)**
A1	6338	11	13	11	9	10	97.95
A2	10	6714	12	9	14	9	97.93
A3	6	12	5948	18	21	11	97.31
A4	11	12	12	6950	62	97	94.79
A5	9	9	8	92	5826	71	94.20
A6	12	11	9	73	80	5077	93.50
P (%)	99.25	99.19	99.10	97.16	96.91	96.25	
F (%)	99.20	99.19	98.98	97.22	96.88	96.37	

Table VIII. Confusion matrix of proposed MS-IB-HHAR *using Mix2 on HHAR **UCI** test set*

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**R (%)**
A1	6352	5	12	6	9	8	99.37
A2	11	6719	7	5	14	12	99.28
A3	8	9	5968	9	11	11	99.20
A4	4	11	7	7055	24	43	98.75
A5	5	10	8	34	5907	51	98.20
A6	2	8	4	29	38	5181	98.46
P (%)	99.53	99.36	99.37	98.84	98.40	97.64	
F (%)	99.45	99.32	99.28	98.80	98.30	98.05	

where TP, FB, TN, and FN represent the True Positive, False Positive, True Negative, and False Negative values respectively.

C. Results

In this part, we present the results of evaluating the proposed MS-IB-HHAR on the six time-series imaging techniques using the test set of HHAR dataset. Table III presents the confusion matrix for ARP images, and it could be noted that A1, A2, and A3 exhibit high precision exceeding 98%, while A4, A5, and A6 exhibit lower precision of 96.14%, 95.04%, and 94.79% respectively. The confusion matrices for both AGASF and AGADF images are presented in Table IV, and Table V, respectively. As with ARP, A1, A2, and A3 realized high precision value over 99%, and the most classification error occurs in the classes of A4, A5, and A6. However, the usage of AGADF images improves the precision of these classes by 1%; while the usage of AGASF images attains 2% improvements on the precision of these classes compared with the usage of ARP. This could be explained by observing that transforming time-series into polar coordinate enable the capture of temporal characteristics of the time-series which is more beneficial for determining activity class.

Additionally, in Table VI, we present the confusion matrix for the MTF images, from which it is evident that the activities of A1, A2, and A3 obtained 97% precision, while the A4, A5, and A6 classes had significant numbers of misclassifications with decreased lower precision of 96.40%, 93.39% and 92.53% respectively. This occurs because modeling time-series associations using the ARP is more efficient than capturing state changes as in the MTF representation.

The confusion matrix of Mix1 images is depicted in Table VII, from which it can be observed that the usage of Mix1 images attains similar performance as achieved by the AGASF representation. This occurs because the features extracted from AGASF have a greater contribution to classification decisions compared with AGADF features. Furthermore, Table VIII displays the confusion matrix calculated from the usage of the Mix2 image. It is obvious that the classification precisions on A4, A5, and A6 classes (98.84 %, 98.40%, 97.64%) are much improved compared with other image representations. However, the precisions of A1, A2, and A3 classes are still larger by 1%-2%. This can be explained by the fact that the combination of features of several image formats enable the acquisition of more information about various activities, resulting in improved performance. It is notable that most of the

Table IX. The performance of the proposed MS-IE-HHAR using different imaging technique on HHAR UCI dataset

Encoding technique	Accuracy (%)	Precision (%)	Recall (%)	F1-measure (%)
Chebyschev [30]	82.62%	78.90%	84.37%	81.54%
Minkowski [30]	84.58%	82.90%	85.84%	84.34%
MS-IE-HHAR+RP	95.59%	93.11%	96.24%	94.65%
MS-IE-HHAR+ARP	97.17%	97.08%	97.12%	97.10%
MS-IE-HHAR+MTF	96.02%	95.89%	95.95%	95.92%
MS-IE-HHAR+D-GASF	94.02%	93.94%	93.88%	93.91%
MS-IE-HHAR+S-GASF	95.42%	94.87%	96.10%	95.48%
MS-IE-HHAR+AGASF	98.06%	98.04%	97.96%	98.00%
MS-IE-HHAR+D-GADF	93.11%	93.14%	93.27%	93.20%
MS-IE-HHAR+S-GADF	94.25%	94.24%	96.34%	95.28%
MS-IE-HHAR+AGADF	97.91%	97.86%	97.85%	97.86%
MS-IE-HHAR+Mix1	98.02%	97.98%	97.98%	97.98%
MS-IE-HHAR+Mix2	98.90%	98.86%	98.88%	98.87%

Table X. Comparison of the results achieved by the proposed MS-IE-HHAR and the recent studies on HHAR UCI dataset.

Study	Accuracy	Precision	Recall	F1-measure
RBM [32]	69.72%	65.11%	61.72%	63.37%
Multi-RBM [33]	81.12%	78.81%	97.35%	87.10%
RF [28]	76.24%	73.34%	75.02%	74.17%
SVM [31]	78.21%	77.11%	76.24%	76.67%
CSA-RF [35]	90.23%	82.34%	83.17%	82.75%
CSA-NN [35]	89.14%	79.38%	80.97%	80.17%
FGA-RF [35]	92.57%	84.14%	86.14%	85.13%
FGA-NN [35]	92.23%	86.27%	90.02%	88.11%
MRP + ResNet [12]	91.16%	85.71%	88.98%	87.31%
FCN [34]	86.89%	83.79%	87.37%	85.54%
ResNet-GASF [30]	95.88%	94.96%	96.01%	95.48%
ResNet-GADF [30]	95.18%	93.55%	94.87%	94.21%
MS-IE-HHAR+ARP	97.17%	97.08%	97.12%	97.10%
MS-IE-HHAR+MTF	96.02%	95.89%	95.95%	95.92%
MS-I*E*-HHAR-AGASF	98.06%	98.04%	97.96%	98.00%
MS-IE-HHAR+AGADF	97.91%	97.86%	97.85%	97.86%
MS-IE-HHAR+Mix1	98.02%	97.98%	97.98%	97.98%
MS-IE-HHAR+Mix2	**98.90%**	**98.86%**	**98.88%**	**98.87%**

Table XI. Comparison of the results achieved by the proposed MS-IE-HHAR and the recent studies on MHEALTH UCI dataset

	ACC+GYR		ACC+MAG		ACC+GYR+MAG	
Study	Accuracy	F1-measure	Accuracy	F1-measure	Accuracy	F1-measure
FEM + SVM [31]	69.13%	66.6%	70.8%	69.84%	70.8%	69.84%
CSA-RF [35]	91.02%	90.13%	90.21%	90.09%	91.12%	90.52%
FGA-RF [35]	93.58%	93.22%	93.17%	93.12%	93.82%	93.43%
MRP+ResNet[12]	92.16%	93.01%	90.97%	91.16%	92.24%	91.84%
FCN [34]	91.22%	90.94%	94.90%	93.94%	96.13 %	95.94%
ResNet-GASF [30]	97.63 %	97.75%	97.52%	97.18%	97.39%	97.26%
ResNet-GADF [30]	97.12%	96.84%	97.29%	97.83%	97.01%	96.97%
MS-IE-HHAR+ARP	98.14%	98.05%	98.27%	98.41%	98.85%	98.11%
MS-IE-HHAR+MTF	96.96%	96.75%	96.43%	96.27%	96.17%	96.58%
MS-I*E*-HHAR-AGASF	99.12%	98.16%	99.08%	99.12%	99.25%	99.24%
MS-IE-HHAR+AGADF	98.53%	98.43%	98.94%	98.09%	98.87%	98.92%
MS-IE-HHAR+Mix1	99.03%	98.82%	99.07%	98.86%	99.11%	99.89%
MS-IE-HHAR+Mix2	99.68%	98.83%	99.81%	99.16%	99.59%	99.34%

encountered errors fall in forecasting stairs up, stairs down, and walking, and this can be explained due to diversity in stairs inclination, or the position of attaching a smartphone to the subject's body.

D. Comparative Analysis

In this section, two comparative experiments were conducted; one for analyzing the importance of introduced encoding techniques, and the other to validate the superiority of the proposed MS-IE-HHAR.

Firstly, we compared the adapted image encoding techniques with each other and with conventional techniques, and the results is presented in Table IX. It can be noted that the introduced ARP technique significantly outperforms the Chebyschev and Minkowski [30] technique, and also it outperforms conventional RP by 2% in accuracy and 3% in F1-measure. It also notable that the MTF encoding shows better performance than the conventional RP, yet shows lower performance than ARP as previously discussed. In addition, we experimented with the proposed MS-IE-HHAR using D-GASF, S-GASF, and AGASF (i.e. D-GASF + S-GASF), the results

Table XII. The contribution of different building blocks to the final performance on HHAR dataset using Different encoding techniques.

	ARP		**AGASF**		**AGADF**		**MTF**		**Mix1**		**Mix2**	
Component	**A**	**F1**	**A**	**F1**	**A**	**F1**	**A**	**F1**	**A**	**F1**	**A**	**F1**
Baseline	90.25%	90.34%	92.82%	92.35%	91.03%	90.05%	89.01%	89.12%	91.37%	91.67%	92.15%	92.24%
Baseline + HME	93.76%	93.83%	94.39%	94.19%	92.93%	93.04%	91.88%	91.82%	94.54%	94.49%	95.39%	95.62%
Baseline + SwA	92.41%	92.54%	94.15%	94.09%	93.33%	92.31%	91.87%	91.43%	93.12%	93.52%	93.87%	94.01%
Baseline + CwA	92.89%	92.14%	93.98%	94.19%	93.84%	92.54%	90.31%	90.54%	93.48%	94.07%	94.39%	94.02%
Baseline+ ISCA	93.35%	93.49%	95.63%	95.57%	94.46%	93.76%	92.45%	92.89%	94.89%	94.97%	95.01%	95.15%
Baseline + HME+ SwA	96.37%	96.45%	96.69%	96.89%	96.28%	96.14%	94.54%	94.32%	97.21%	97.37%	97.11%	97.37%
Baseline + HME+ CwA	95.95%	95.91%	96.96%	96.21%	95.84%	96.14%	94.02%	93.82%	97.14%	97.02%	97.18%	97.91%
MS-IE-HHAR	97.17%	97.10%	98.06%	98.00%	97.91%	97.86%	96.02%	95.92%	98.02%	97.98%	98.90%	98.87%

show that employing the S-GASF improves the performance by 1% in accuracy and 2% in F1-measure over the D-GASF. It is also noted that the AGASF realizes 3% improvements over S-GASF on all measures. Similar observations can be made when experimenting with MS-IE-HHAR using D-GADF, S-GADF, and AGADF. Nevertheless, the AGADF image shows 1% lower performance than the AGASF. This further demonstrates the difference between normalizing dynamic and static activities in polar coordinates.

The results attained from Mix1 encoding show 1% accuracy improvements over AGADF. However, they did not show any improvement over AGASF encoding, which further validates the effectiveness of AGASF encoding in modeling time-series characteristics. Furthermore, the results attained by Mix2 encoding show robust performance (Accuracy: 98.90%, Precision: 98.86%, Recall: 98.88%, F1-measure: 98.87%) exceeding all of the other encoding techniques investigated. This demonstrates the effectiveness of combining the MTF features with AGASF, and AGADF, which in turn enables the classifier to use state change information (from MTF) and temporal encoded information (from GASF).

Secondly, an additional comparative experiment is performed to compare the proposed approach with the recently introduced models for HRAR, and the corresponding results are presented in Table X. it could be seen that the proposed MS-IE-HHAR outperformed all of the current approaches under different encoding techniques. For example, the MS-IE-HHAR outperformed the time-series based studies [28], [31], [35] with 5%-7% in accuracy and 10%-13% in F1-measure. Compared with ResNet [12], the proposed MS-IE-HHAR shows 6% accuracy improvements and 10% F1-measure improvements. Compared with FCN [34], MS-IE-HHAR shows 10% improvements in accuracy and F1-measure. Further, the proposed MS-IE-HHAR achieves 2%-3% improvements over ResNet [30] in both accuracy and F1-measure. This observation demonstrates the efficiency of the proposed MS-IE-HHAR model and its constituent modules of HME and ISCA.

To further validate the efficiency and the effectiveness of the proposed MS-IE-HHAR, additional comparative evaluation experiments were performed on the MHEALTH dataset. In this experiment, we assessed model performance on different combinations of sensors, namely: accelerometer and gyroscope, accelerometer and magnetometer, and all of these sensors. In view of this, Table XI shows the yielded results from evaluating the models in the three scenarios outlined above . It can be seen that the MS-IE-HHAR shows robust performance on a different combinations of sensory data. For instance, it outperforms ResNet [12], and FCN [34] with 5%-10% in accuracy in all data combination scenarios, and realize 1-2% improvements over the ResNet [12]. These results provide clear evidence of the superiority of the proposed model as with the HHAR UCI dataset.

E. Ablation Study

For convenience, in our experiments, we select Res2Net [47] as a baseline architecture, and then perform several experiments on the HHAR dataset to determine the contribution of each building block to the final model performance using different encoding techniques—the results obtained from these experiments are presented in Table XII. It can be noted that redesigning the baseline using the proposed HME module result in 2%-3% improvements in accuracy and F1-measure. This demonstrates that the multi-scale feature extraction implemented by the HME module is vitally important for fine-tuning the performance of HAR from imaged time-series.

In addition, integrating the SwA block between the residual block of the baseline architecture has been shown to increase the performance by 1%-2% on types of images. A similar observation was noticed when applying the CwA block in the same manner. Further, the performance improvements are observed when applying these blocks to the Baseline + HME architecture. This validates that attending to spatial correlations and channel interdependency is of great importance for classifying human activities within imaged time-series. Moreover, the parallel combination of SwA and CwA in ISCA module brings the advantages of both of them with a 3%-4% increase in performance over the baseline.

V. Limitations and Future work

Despite of demonstrated efficiency of the proposed model, it still has some limitations that might be addressed in future studies. First, the proposed model is trained in a supervised manner, so it could not benefit from the large number of unlabeled samples that are available in many IoT environments. Semi-supervised models (i.e. Generative Adversarial Network, Self-ensembling, etc.) can offer an efficient solution for this, which we intended to investigate in near future. Second, the proposed model could not be generalized for other HAR data such as WIFI signals and radar data streams. Accordingly, we intended to investigate HAR task time-series emanating from WIFI devices and radar devices. Third, to prepare for future business challenges we integrate the proposed framework in the Economy of Things (EoT) environment by leveraging distributed deep learning competencies for monetization of instantaneous HAR data generated from different IoT devices. Fourth, the recent advances in blockchain technologies make the IoT research community conceive new challenges for security control and data monetization. Nevertheless, developing a blockchain-enabled IoT system is still challenging owing to the heterogeneous nature of blockchain platforms and the lack of guidelines on how to interface existing components in the IoT ecosystem with the emerging blockchain technology. Thus, we propose to extend and deploy the proposed model to address these challenges.

VI. Conclusion

This paper introduced a novel framework, called MS-IE-HHAR, for fine-grained human activity recognition from heterogeneous sensors (usually embedded in a variety of IoT devices) aiming to enable monetization of data of consumers of HAR applications within the IoT ecosystem. First, we redesign three techniques for encoding HAR sensory data into a multi-channel image representation. An innovative deep learning model (i.e. MS-IE-HHAR) is applied to classify the activities based on the generated images. An HME is proposed to enable efficient multi-scale feature extraction from different receptive fields. The ISCA module is proposed to capture spatial correlations and channel interdependency information which is proven to improve the classification performance of imaged time-series. Furthermore, for minimizing the generalization error of our model, we adopted a hybrid loss function based on cross-entropy loss and pairwise Gaussian loss. The results proved the feasibility of our proposed approaches in recognizing activity image-encoded time series from heterogeneous inertial sensors.

References

[1] T. Huynh-The, C. Hua, N. A. Tu and D. Kim, "Physical Activity Recognition with Statistical-Deep Fusion Model using Multiple Sensory Data for Smart Health," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2020.3013272.

[2] N. Rashid, M. Dautta, P. Tseng and M. A. A. Faruque, "HEAR: Fog-enabled Energy Aware Online Human Eating Activity Recognition," in *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2020.3008842.

[3] X. Niu, L. Xie, J. Wang, H. Chen, D. Liu and R. Chen, "AtLAS: An Activity-based Indoor Localization and Semantic Labeling Mechanism for Residences," in *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2020.3004496.

[4] X. Zhou, W. Liang, K. I. Wang, H. Wang, L. T. Yang and Q. Jin, "Deep-Learning-Enhanced Human Activity Recognition for Internet of Healthcare Things," in IEEE Internet of Things Journal, vol. 7, no. 7, pp. 6429-6438, July 2020, doi: 10.1109/JIOT.2020.2985082.

[5] T. Li, S. Fong, K. K. Wong, Y. Wu, X.-s. Yang, and X. Li, "Fusing Wearable and Remote Sensing Data Streams by Fast Incremental Learning with Swarm Decision Table for Human Activity Recognition," Information Fusion, 2020.

[6] J. Zhang *et al*., "Data Augmentation and Dense-LSTM for Human Activity Recognition using WiFi Signal," in *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2020.3026732.

[7] O. Barut, L. Zhou and Y. Luo, "Multitask LSTM Model for Human Activity Recognition and Intensity Estimation Using Wearable Sensor Data," in *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8760-8768, Sept. 2020, doi: 10.1109/JIOT.2020.2996578.

[8] M. Chen, Y. Li, X. Luo, W. Wang, L. Wang and W. Zhao, "A Novel Human Activity Recognition Scheme for Smart Health Using Multilayer Extreme Learning Machine," in IEEE Internet of Things Journal, vol. 6, no. 2, pp. 1410-1418, April 2019, doi: 10.1109/JIOT.2018.2856241.

[9] M J. Lu, X. Zheng, M. Sheng, J. Jin and S. Yu, "Efficient human activity recognition using a single wearable sensor," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2020.2995940.

[10] C.-L. Yang, Z.-X. Chen, and C.-Y. Yang, "Sensor Classification Using Convolutional Neural Network by Encoding Multivariate Time Series as Two-Dimensional Colored Images," Sensors, vol. 20, p. 168, 2020.

[11] V. M. Souza, D. F. Silva, and G. E. Batista, "Extracting texture features for time series classification," in 2014 22nd International Conference on Pattern Recognition, 2014, pp. 1425-1430.

[12] J. Lu and K.-Y. Tong, "Robust Single Accelerometer-Based Activity Recognition Using Modified Recurrence Plot," IEEE Sensors Journal, vol. 19, pp. 6317-6324, 2019.

[13] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," in Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[14] A. Estebsari and R. Rajabi, "Single Residential Load Forecasting Using Deep Learning and Image Encoding Techniques," Electronics, vol. 9, p. 68, 2020.

[15] F. Setiawan, B. N. Yahya, and S.-L. Lee, "Deep activity recognition on imaging sensor data," Electronics Letters, vol. 55, pp. 928-931, 2019.

[16] W. Wu and Y. Zhang, "Activity Recognition from Mobile Phone using Deep CNN," in 2019 Chinese Control Conference (CCC), 2019, pp. 7786-7790.

[17] D. F. Silva, V. M. De Souza, and G. E. Batista, "Time series classification using compression distance of recurrence plots," in 2013 IEEE 13th International Conference on Data Mining, 2013, pp. 687-696.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision. 2015," arXiv preprint arXiv:1512.00567, 2015.

[20] Y. Kong and Y. Fu, "Max-margin heterogeneous information machine for RGB-D action recognition," International Journal of Computer Vision, vol. 123, pp. 350-371, 2017.

[21] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, "Cooperative training of deep aggregation networks for RGB-D action recognition," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[22] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in neural information processing systems, 2014, pp. 568-576.

[23] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," arXiv preprint arXiv:1803.02155, 2018.

[24] Y. Qin, G. Liu, Z. Li, C. Yan, and C. Jiang, "Pairwise Gaussian loss for convolutional neural networks," IEEE Transactions on Industrial Informatics, 2020.

[25] N. McLaughlin, J. Martinez-del-Rincon, and P. Miller, "3-D Human Pose Estimation Using Iterative Conditional Squeeze and Excitation Networks," IEEE Transactions on Cybernetics, 2020.

[26] A. Guha Roy, N. Navab, and C. Wachinger, "Recalibrating Fully Convolutional Networks with Spatial and Channel'Squeeze & Excitation'Blocks," arXiv preprint arXiv:1808.08127, 2018.

[27] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel “squeeze and excitation” blocks," IEEE transactions on medical imaging, vol. 38, pp. 540-549, 2018.

[28] H. Blunck, S. Bhattacharya, A. Stisen, T. S. Prentow, M. B. Kjærgaard, A. Dey, et al., "Activity recognition on smart devices: Dealing with diversity in the wild," GetMobile: Mobile Computing and Communications, vol. 20, pp. 34-38, 2016.

[29] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J. A. Holgado-Terriza, et al., "Design, implementation and validation of a novel open framework for agile development of mobile health applications," Biomedical engineering online, vol. 14, p. S6, 2015.

[30] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K.-K. R. Choo, "Imaging and fusing time series for wearable sensor based human activity recognition," Information Fusion, vol. 53, pp. 80-87, 2020.

[31] N. Y. Hammerla, R. Kirkham, P. Andras, and T. Ploetz, "On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution," in Proceedings of the 2013 international symposium on wearable computers, 2013, pp. 65-68.

[32] S. Bhattacharya and N. D. Lane, "From smart to deep: Robust activity recognition on smartwatches using deep learning," in 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), 2016, pp. 1-6.

[33] V. Radu, N. D. Lane, S. Bhattacharya, C. Mascolo, M. K. Marina, and F. Kawsar, "Towards multimodal deep learning for activity recognition on mobile devices," in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, 2016, pp. 185-188.

[34] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in 2017 International joint conference on neural networks (IJCNN), 2017, pp. 1578-1585.

[35] M. Ehatisham-Ul-Haq, M. A. Azam, Y. Amin, and U. Naeem, "C2FHAR: Coarse-to-Fine Human Activity Recognition With Behavioral Context Modeling Using Smart Inertial Sensors," IEEE Access, vol. 8, pp. 7731-7747, 2020.

[36] F. Firouzi, K. Chakrabarty, and S. Nassif, *Intelligent Internet of Things: From Device to Fog and Cloud*: Springer, 2020.

[37] F. Firouzi, B. Farahani, and M. N. Bojnordi, "The Smart “Things” in IoT," in *Intelligent Internet of Things*, ed: Springer, 2020, pp. 51-95.

[38] F. Firouzi, B. Farahani, M. Weinberger, G. DePace, and F. S. Aliee, "IoT Fundamentals: Definitions, Architectures, Challenges, and Promises," in *Intelligent Internet of Things*, ed: Springer, 2020, pp. 3-50.

[39] F. Firouzi, B. Farahani, F. Ye, and M. Barzegari, "Machine Learning for IoT," in *Intelligent Internet of Things*, ed: Springer, 2020, pp. 243-313.

[40] F. Firouzi and B. Farahani, "Architecting IoT Cloud," in *Intelligent Internet of Things*, ed: Springer, 2020, pp. 173-241.

[41] D. Purwanto, R. R. A. Pramono, Y.-T. Chen, and W.-H. Fang, "Three-stream network with bidirectional self-attention for action recognition in extreme low resolution videos," *IEEE Signal Processing Letters,* vol. 26, pp. 1187-1191, 2019.

[42] D. Purwanto, R. Renanda Adhi Pramono, Y.-T. Chen, and W.-H. Fang, "Extreme low resolution action recognition with spatial-temporal multi-head self-attention and knowledge distillation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0-0.

[43] S. Cho, M. Maqbool, F. Liu, and H. Foroosh, "Self-Attention Network for Skeleton-based Human Action Recognition," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 635-644.

[44] Q. Liu, X. Che, and M. Bie, "R-STAN: Residual spatial-temporal attention network for action recognition," *IEEE Access,* vol. 7, pp. 82246-82255, 2019.

[45] J. Lei, Y. Jia, B. Peng, and Q. Huang, "Channel-wise Temporal Attention Network for Video Action Recognition," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 562-567.

[46] A. Diba, M. Fayyaz, V. Sharma, M. Mahdi Arzani, R. Yousefzadeh, J. Gall, *et al.*, "Spatio-temporal channel correlation networks for action classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 284-299.

[47] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[48] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848-6856