Interpretable Parallel Recurrent Neural Networks with Convolutional Attentions for Multi-Modality Activity Modeling

Kaixuan Chen*, Lina Yao*, Xianzhi Wang*, Dalin Zhang*, Tao Gu†, Zhiwen Yu‡ and Zheng Yang§

*School of Computer Science and Engineering, UNSW Sydney, Australia
†School of Information Technology, RMIT University, Australia
‡School of Computer Science, Northwestern Polytechnical University, China
§School of Software, Tsinghua University China

Abstract—Multimodal features play a key role in wearable sensor based human activity recognition (HAR). Selecting the most salient features adaptively is a promising way to maximize the effectiveness of multimodal sensor data. In this regard, we propose a "collect fully and select wisely" principle as well as an interpretable parallel recurrent model with convolutional attentions to improve the recognition performance. We first collect modality features and the relations between each pair of features to generate activity frames, and then introduce an attention mechanism to select the most prominent regions from activity frames precisely. The selected frames not only maximize the utilization of valid features but also reduce the number of features to be computed effectively. We further analyze the accuracy and interpretability of the proposed model based on extensive experiments. The results show that our model achieves competitive performance on two benchmarked datasets and works well in real life scenarios.

Index Terms—HAR, attention, deep learning, wearable sensors

I. INTRODUCTION

HAR plays a key role in several research fields. It has gained broad attention due to the increasing popularity of ubiquitous environments, especially in health care and surveillance domains [1].

Despite a large number of sensor-based recognition solutions proposed over the decade, we discover several limitations. First, there is still a lack of comprehensive model representation to sensor signals in a way that different activities can be distinguished in a more expressive and effective ways. With the recent advances in deep neural networks and the notable performance achieved by these methods in the community of HAR [2], Convolutional Neural Network (CNN) appears to be a promising candidate for building such models. However, while CNN does well in capturing spatial relationships of features, it focuses merely on the features covered by the convolutional kernels but overlooks the correlation among non-adjacent features [3]. Considering that most of the data collected by wearable sensors such as accelerometers and gyroscopes are tri-axis, in this paper, we transform sensor signals into new activity frames which not only capture the relationships between each pair of tri-axis signals but also contain the relations between each pair of single signals. The experiments show that our new representation is far more discriminative than traditional representations.

Second, the demerits of interperson variability and interclass similarity can greatly reduce system performance [4]. Interperson variability comes from the fact that the same activity can be performed differently by different people, and interclass similarity results from the similarity in the behavior patterns of different activities like walking and running. Both the above issues require the classifier to be task dependent, i.e., it should automatically extract the salient information indicative of the true activity and ignore the interclass similarity. To this end, we propose an attention based model, which is directly related to the HAR tasks, to address the problems of interperson variability and interclass similarity.

Attention is originally a concept in biology and psychology that implies focusing the power of noticing or thinking on something special to achieve better cognitive processes. The attention mechanisms have several advantages, the first being task dependence. Intuitively, the motion of different body parts has varied contributions to different activities [5]. For example, jumping mostly involves legs while running is related to both arms and legs. More specifically, recognizing the patterns of walking depends more on the acceleration of legs while distinguishing sitting from lying would rely more on the orientation. In this paper, we separate the data related to each body part to different modals, namely acceleration, angular velocity and magnetism. The attention mechanisms ensure that the system only focuses on the most contributing data and ignores the irrelevant sensors or modalities.

The second advantage of the attention mechanisms is that it opens the black box of deep neural networks to a certain degree. While the inner mechanisms of neural networks remain implicit, interpretable neural network is becoming another trend in the machine learning and data mining fields. Taking CNN for example, when CNN recognizes a dog from an image, we tend to explicitly know that one filter distinguishes the dog head and another filter identifies the dog paw. Back to activity recognition, the attention model not only provides the specific body parts it focuses on but also highlights the most contributing sensors and modals to distinguish diverse

activities. The salient sensor data can be inferred from the glimpse patch (to be detailed in Section II-B2).

The third advantage is that it reduces the computational cost significantly. Usually, the dimension of the features expands as we extract the full spatial relationships among sensors, and the cost increases with the increase of input data dimension. Most existing models process the entire data every time, resulting in high computational cost. Some works [6], [7] aim to limit the input dimension using techniques such as dimensionality reduction and feature selection. However, feature processing comes with information loss, leading to a new trade-off problem between accuracy and cost. Inspired by human attention, our proposed method focuses on only one small patch of the data each time and goes to the next patch when necessary. This method considerably reduces computational cost as well as information loss.

In this paper, we tackle the HAR problems by transforming wearable sensor data into activity frames and deploying an interpretable parallel recurrent model with convolutional attentions, including one attention based LSTM and one activity frame based LSTM, to recognize activities. The main contributions of this work are summarized as follows:

- We transform the tri-axis sensor data into activity frames to extract the full relationships between data pairs. This enables the CNN to cover all features without overlooking any relationships between data pairs.
- We propose a parallel recurrent model including one attention based LSTM and one activity frame based LSTM to recognize activities. Firstly, the system focuses on only a small patch of the activity frame that contains the most salient information to avoid unnecessary cost on less important areas, by leveraging the attention based LSTM and combining reinforcement learning. Secondly, we deploy a activity frame based LSTM to exploit spatial and temporal information in time-series signals and capture the dynamics of the sensor data.
- We examine our model on two public benchmarked datasets, PAMAP2 [8] and MHEALTH [9] and perform extensive comparison with other methods, as well re-examine our approach on a new dataset collected in the real world named MARS. The experimental results show that our proposed model consistently outperforms a series of baselines and state-of-the-arts.

II. OUR MODEL

To fully collect effective information and wisely select salient features, our model contains two parts: (a) feature extraction to firstly transform wearable sensor data into 2-D matrices. (b) an interpretable parallel recurrent model with convolutional attentions including one attention based LSTM and one activity frame based LSTM for activity recognition. Our attention based LSTM simulates the procedures of human brains processing visual information within several glimpses. The other one is activity frame based LSTM. Owing to the facts that the activity recognition largely depends on temporal information and that activity frames naturally capture serial

relations, activity frame based model is more suitable for our scenarios.

The above process is presented as a three-dimensional model in Figure 1, where the time step t and frame f represent the attention based LSTM and the activity frame based LSTM in our method, respectively.
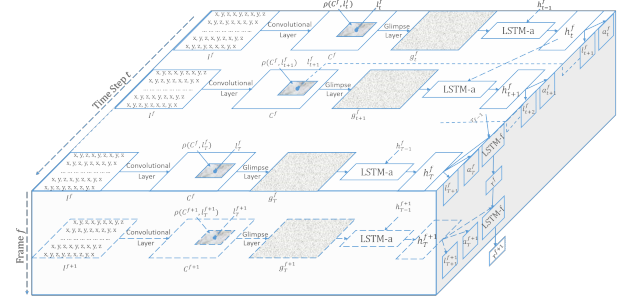

Fig. 1. Work-flow of the Proposed Approach. Dashed arrows indicate the time step t for attention based LSTM and the frame f for activity frame based LSTM, respectively. For each time step t, the input frame goes through a convolutional network to obtain a higher-level representation C^f. We extract a retina region $\rho(C^f, l_t^f)$ at location l_t^f, which is decided by the last time step $t-1$. $\rho(C^f, l_t^f)$ next goes through a glimpse layer to get the glimpse g_t^f as input of the attention based LSTM-a which decides the action a_t^f and the next location l_{t+1}^f. For the activity frame based LSTM-f takes the last action of each frame a_f^T as input and outputs the final prediction.

A. Input Representation

As we transform the wearable sensor data into activity frames, the data are represented as three-dimensional vectors. Each sample $(\mathbf{x}, \mathbf{y})$ of the model consists of a 3-d vector $\mathbf{x}$ and the activity label $\mathbf{y}$. Suppose X, Y, F denote activity frames' width, height, and number of frames, and C represents the number of activity classes, we have: $\mathbf{x} \in R^{X \times Y \times F}$ and $\mathbf{y} \in [1, ..., C]$.

1) Activity Frame: There already exist some previous works that combine multimodal wearable sensor data for HAR in feature level. For example, Kunze et al. [10] concatenate acceleration and angular velocity into one vector and [11]–[13] combine acceleration and other modalities including microphone and GPS data. However, these works overlook the relations among sensors which are important to activity recognition. A popular method for extracting spatial relations is deep learning methods like CNN. Although CNN is proven to perform well in HAR [14], [15], the accuracy is still not that satisfactory. In fact, CNN is originally proposed for images where each pixel is only related to its adjacent pixels and this small area can be easily covered by a kernel patch of a convolutional layer. However, it is still challenging to transform features to extract relations between each signal and the related signals for HAR. In many cases of HAR [16], the sensor data are arranged according to the physical connection of human body parts. For example, the sensor data of hands should be adjacent to the data of shoulders and the data of shoulders should be adjacent to the data of the waist, which should be followed by the data of hips, legs, and feet.

Nevertheless, in the real world, activities always depend on more than one body part. For instance, running relies on the cooperation of arms and legs. In addition, the common Inertial Measurement Unit in wearable devices usually includes a tri-axis accelerometer, a tri-axis gyroscope, and a tri-axis magnetometer, and the degree to which these sensors contribute to different activities are various. This makes it even more important to find a representative transformation to extract the relationships between each pair of tri-axis sensor signals (e.g. acceleration and angular velocity) and each pair of single signals (e.g. the first dimension of acceleration and the second dimension of angular velocity).
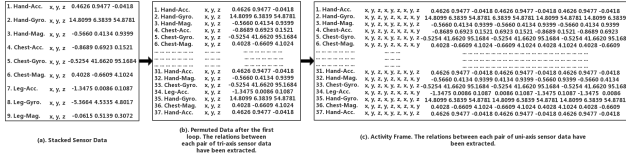

Fig. 2. Transformation from sequences to frames

Figure 2 shows the transformation process into activity frames. Each figure is comprised of four parts: sequence number, sensor location (hand, chest, leg) and modality (acceleration, angular velocity...), notations (x, y, z), and real data examples. Algorithm 1 further illustrates the transmigration from sequences to images. First, raw signals are stacked row-by-row as shown in Figure 2 (a). After being permuted in the first loop (Line 6-18 in Algorithm 1), each tri-axis sensor data has a chance to be adjacent to each of the other sensor data as shown in Figure 2 (b). For example, supposing $N_r = 9$, then the final S_p is $[1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 3, 5, 7, 9, 2, 4, 6, 8, 1, 4, 7, 1, 5, 8, 2, 5, 9, 3, 6, 9, 4, 8, 3, 7, 2, 6, 1]$. Since we still need to extract the relationships between each pair of single sensor signals, the second loop (Line 19-25 in Algorithm 1) ensures that each single signal has a chance to be adjacent to each of the other signals as Figure 2 (c) shows. So far we have extracted the relationships between each pair of single sensor signals.

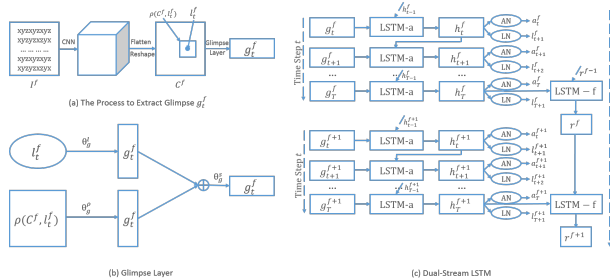*B. Parallel Recurrent Model with Convolutional Attentions*

Fig. 3. Flattened Model. (a) Extracting glimpse g_t^f from the input activity frame, including a CNN, flattening and reshaping, and a glimpse layer. (b) The detailed description of the glimpse layer which combines the location l_t^f and the retina region $\rho(C^f, l_t^f)$. (c) Parallel recurrent procedure containing attention based LSTM-a and activity frame based LSTM-f.

ALGORITHM 1: Transformation from Sequences to Activity Frames

Input: Stacked raw signals. Each row is a tri-axis data of a accelerometer, gyroscope or a magnetometer which can be denoted as x, y, z. As shown in Figure 2 (a), each row has a sequence number. Here the number of rows $N_r = 9$ as an example.

Output: The activity frame I_A which is a 2-D array

1: $i = 1$;
2: $j = i + 1$;
3: permutation sequence $S_p = [0]$;
4: adjacent pair set $S_a p = \emptyset$;
5: activity frame I_A = the first row of stacked signals
6: **while** $i \neq j$ **do**
7:   **if** $j > N_r$ **then**
8:     $j = 1$;
9:   **else if** $(i, j) \notin S_a p$ and $(j, i) \notin S_a p$ **then**
10:     add (i, j) to $S_a p$;
11:     add j to S_p;
12:     add the j-th row of input data to I_A;
13:     $i = j$;
14:     $j = i + 1$;
15:   **else**
16:     $j = j + 1$
17:   **end if**
18: **end while**
19: **for** each row of I_A **do**
20:   **if** the sequence number of this row is odd **then**
21:     this row is extended as $'x, y, z, x, y, z, x, y, z'$
22:   **else**
23:     this row is extended as $'x, y, z, y, z, x, z, x, y'$
24:   **end if**
25: **end for**
26: **return** I_A

We propose an interpretable parallel recurrent model with convolutional attentions that incorporates both attention and temporal information to analyze the activity frames. Figure 3 shows the structure of this model, where the activity frame based LSTM recurrent model leverages the temporal information of sensor data and the attention based LSTM solves the HAR problem.

Since human body parts contribute differently in recognizing various activities, we need to guarantee that the system only focuses on the most relevant and contributing parts and data. Some previous works [17], [18] leverages attention for image classification. However, analyzing activity sensor data in this work can be more challenging because sensor data lacks meaningful characteristics compared with image data. Therefore, we combine attention mechanism with CNN and RNN to automatically extract the most salient modality-specific features, further convert the information to higher-level representation and combine the spatial information with temporal information. Since the activity frames fully extract

the relationships among all feature pairs, only a part of them is salient to each certain activity. Therefore, it is natural to introduce attention mechanisms facilitating to mine effective information and minimize the negative impacts of undesirable information. To the best of our knowledge, our method is the first one to leverage the attention model to tackle the activity recognition problems.

Figure 3 shows a flattened model, which better interprets the model. Our model is comprised of a convolutional network, a glimpse network, a recurrent attention unit, and a recurrent activity frame unit that we will introduce in the followings.

1) Convolutional Network: To derive an effective representation of features, we further transform activity frames into convolutional activity frames. Compared with a convolutional auto-encoder [19], we prefer to train the model end-to-end and omit the pretraining process, as shown in Figure 3. Each activity frame I^f (f denoted the f_{th} frame) is transformed into a three-dimensional cube, the height of which depends on the number of channels of the convolutional network. The convolutional network has two convolutional layers that learn filters which activate when it detects some specific types of features at some spatial position in the input. The output is further processed by a ReLU layer and a max pooling layer. The former applies the non-saturating activation function $relu(\nu) = max(\nu, 0)$ to increase the nonlinear properties of both the decision function and the overall network without affecting the receptive fields of the convolution layer. The latter partitions the input image into a set of non-overlapping rectangles and outputs the maximum for each such sub-region to omit the less important features.

To obtain new convolutional activity frames, the cubes are flattened and reshaped in the same size of original activity frames by a fully connected layer. After the convolutional layer, the input frame I^f is encoded to be C^f.

2) Glimpse Network: The first part after the convolutional network is a glimpse network. The glimpse network not only avoids the system processing the whole data in the entirety at a time but also maximally eliminates the information loss. In our model, each frame will be "understood" within T glimpses. For the transformed frame C^f, at each time step t, we simulate the process of how the human eyes work. Our model first extracts a retina region denoted by $\rho(C^f, l_t^f)$ from the input data at the location l_t^f with a retina. The retina image encodes the region around l_t^f with high resolution but uses a progressively lower resolution for points further from l_t^f. This has been proved an effective method to remove noises and avoid information loss in [20].

In the human visual system, the retina image is converted into electric signals that are relayed to the brain via the optic nerves. Likewise, in our model, the retina image is converted into a glimpse g_t^f as Figure 3 shows. The retina image $\rho(C^f, l_t^f)$ and the location l_t^f are linearly transformed independently with two linear layers parameterized by θ_g^ρ and θ_g^l, respectively. Next, the summation of these two parts is further transformed with another linear layer parameterized

by θ_g^s and a rectified linear unit. The whole process can be summarized as the following equation:

$$g_t^f = f_g(\rho(C^f, l_t^f), l_t^f; \theta_g^\rho, \theta_g^l, \theta_g^s)$$
$$= relu(Linear(Linear(\rho(C^f, l_t^f)) + Linear(l_t^f)))$$

where $Linear(\bullet)$ denotes a linear transformation. Therefore, g_t^f contains information from both "what" ($\rho(C^f, l_t^f)$) and "where" (l_t^f).

3) Recurrent Attention Unit: We use the recurrent neural networks as the core to process data step by step within several glimpses and introduce an attention mechanism to ensure the system only focuses on the most relevant sensors/modals and the most contributing data. The glimpses at time steps of the attention based LSTM help visualize the contribution of sensors deployed at different body parts, thus achieving better interpretability of our model.

As Figure 3 shows, the basic structure of the recurrent attention unit is an LSTM-a (attention based LSTM). At each time step t, the LSTM-a receives the glimpse g_t^f and the previous hidden state h_{t-1}^f as the inputs parameterized by θ_h. Meanwhile, it outputs the current hidden state h_t^f according to the equation:

$$h_t^f = f_g(h_{t-1}^f, g_t^f; \theta_h) \tag{1}$$

The recurrent attention model also contains two sub-networks: the location network and the action network. These two sub-networks receive the hidden state h_t^f as the input to decide the next glimpse location l_{t+1}^f and the current action a_t^f. The current action not only determines the activity label $\hat{\mathbf{y}}$ but also affects the environment in some cases while the location network outputs the location at time $t + 1$ stochastically according to the location policy defined by a Gaussian distribution stochastic process, parameterized by the location network $f(h_t^f; \theta_t)$. As it decides the next region to "look at", the location network is the principal component of the recurrent attention unit.

$$l_{t+1}^f \sim P(\cdot \mid f_l(h_t^f; \theta_l)) \tag{2}$$

Similarly, the action network outputs the corresponding action at time t and predicts the activity label given the hidden state h_t^f. The action a_t^f obeys the distribution parameterized by $f(h_t^f; \theta_a)$. Owing to its prediction function, the network uses a softmax formulation:

$$a_t^f = f_a(h_t^f; \theta_a) = softmax(Linear(h_t^f)) \tag{3}$$

4) Recurrent Activity Frame Unit: Activity recognition heavily relies on the temporal information. Therefore, besides the single activity frames used by the aforementioned process, we additionally leverage activity frames via a recurrent activity frame unit. As the hidden layer h_t^f of the core LSTM-a contributes to predicting the action a_t^f and deciding the next glimpse location l_{t+1}^f. For this reason, we believe the hidden state is discriminative enough to make the final prediction

for the whole system. In particular, we design an LSTM-f (activity frame based LSTM) to combine the hidden states of all the frames at the last time step T to predict the activity label and to preserve the efficiency. Given the hidden state of the last frame, the hidden state of each frame $r^f = f_r(h_T^f, r^{f-1}; \theta_r)$, parameterized by θ_r.

C. Training and Optimization

Our proposed model depends on the parameters of every components, including the glimpse network, the recurrent attention network, the two sub-networks, and the activity frame based LSTM, $\Theta = \theta_g, \theta_h, \theta_a, \theta_l, \theta_r$. Both the action network and the frame-based recurrent network are based on classification methods. Therefore, their parameters, θ_a and θ_r, can be trained by optimizing the cross-entropy loss and the backpropagation. However, the location network should be able to select a sequence of salient regions from activity frames adaptively. Since this network is non-differentiable owing to its stochasticity and the problem can also be regarded as a control problem to settle the attention region at the next step, it can be trained by reinforcement methods to learn the optimal policies.

We simply introduce some definitions of reinforcement learning based on our case.

- Agent: the brain to make decisions, which is the location network in our case.
- Environment: the unknown world that may affect the agent's decision or may be influenced by the agent.
- Reward: the feedback from the environment to evaluate the action. In our case, for each frame, the model gives a prediction $\hat{\mathbf{y}} = a_t$ and receives a reward r_t as a feedback for the future correction of the prediction after each time step t. Suppose T denotes the number of steps in our attention based LSTM. $r_t = 1$ if $\hat{\mathbf{y}} = \mathbf{y}$ after T steps and 0 otherwise. The target of the optimization is to maximize $R = \sum_{t=1}^{T} r_t$.
- Policy: the projection from states to actions, denoted by $\pi(a \mid s) = P[A_t = a \mid S_t = s]$. To maximize the reward R, we learn an optimal policy $\pi(l_t, a_t | s_{1:t}; \Theta)$ to map the attention sequence $s_{1:t}$ to a distribution over actions for the current time step, where the policy π is decided by Θ of the recurrent attention model.

Based on the above discussion, we deploy a Partially Observable Markov Decision Process (POMDP) to solve the training and optimization problems, for which the true state of the environment is unobserved. Let $s_{1:t} = \mathbf{x}_1, l_1, a_1; ... \mathbf{x}_t, l_t, a_t$ be the sequence of the input, location and action pairs. This sequence, called an attention sequence, shows the order of the regions our attention focuses on.

To sum up, in our case, the location network is formulated as a random stochastic process (the Gaussian distribution) parameterized by Θ. Each time after the location selection, the prediction a is evaluated to back feed a reward for conducting the backpropagation training process. The process is also defined as policy gradient. Our goal is to maximize the simulated rewards using gradient.

Generally, for sample x with its reward $f(x)$ and the probability $p(x)$, we have:

$$E_x[f(x)] = \sum_x p(x)f(x) \qquad (4)$$

so that the gradient can be calculated according to the REINFORCE rule [21]:

$$
\begin{aligned}
\nabla_\theta E_x[f(x)] &= \nabla_\theta \sum_x p(x)f(x) \\
&= \sum_x \nabla_\theta p(x)f(x) \\
&= \sum_x p(x)\frac{\nabla_\theta p(x)}{p(x)}f(x) \\
&= \sum_x p(x)\nabla_\theta log p(x)f(x) \\
&= E_x[f(x)\nabla_\theta log p(x)] \qquad (5)
\end{aligned}
$$

In our case, given the reward R and the attention sequence $s_{1:T}$, the reward function to be maximized is as follows:

$$J(\Theta) = \mathbb{E}_{p(s_{1:T};\Theta)}\left[\sum_{t=1}^{T} r_t\right] = \mathbb{E}_{p(s_{1:T};\Theta)}[R] \qquad (6)$$

By considering the training problem as a POMDP, a sample approximation to the gradient is calculated as follows:

$$\nabla_\Theta J = \sum_{t=1}^{T} \mathbb{E}_{p(s_{1:T};\Theta)}[\nabla_\Theta log \pi(\mathbf{y}|s_{1:t};\Theta)R] \qquad (7)$$

where i denotes the ith training sample, $\mathbf{y}^{(i}$ is the correct label for the ith sample, and $\nabla_\Theta log \pi(\mathbf{y}^{(i)}|s_{1:t}^i;\Theta)$ is the gradient of LSTM-a calculated by backpropagation.

We use Monte Carlo sampling which utilizes randomness to yield results that might be deterministic theoretically. Supposing M is the number of Monte Carlo sampling copies, we duplicate the same convolutional activity frames for M times and average them as the prediction results to overcome the randomness in the network, where the M duplication generates M subtly different results owing to the stochasticity, so we have:

$$
\begin{aligned}
\nabla_\Theta J &= \sum_{t=1}^{T} \mathbb{E}_{p(s_{1:t};\Theta)}[\nabla_\Theta log \pi(\mathbf{y}|s_{1:t};\Theta)R] \\
&\approx \frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{T} \nabla_\Theta log \pi(\mathbf{y}^{(i)}|s_{1:t}^i;\Theta)R^{(i)}
\end{aligned}
\qquad (8)
$$

Therefore, although the best attention sequences are unknown, our proposed model can learn the optimal policy in the light of the reward.

III. EXPERIMENTS

In this section, we present the systematically validation of our proposed method via experiments on on two public datasets and one real-world dataset collected by ourselves.

A. Datasets and Experimental Settings

We evaluate the proposed method on two public bench-marked activity recognition datasets, PAMAP2 dataset and MHEALTH dataset and the real-world dataset MARS which is collected by ourselves. These public datasets are the latest available wearable sensor-based datasets with complete annotation and have been widely used in the activity recognition research community.

PAMAP2 [8]: The PAMAP2 dataset contains 9 participants performing 12 daily living activities including both basic actions and sportive exercises. The activity sensory data is collected from 3 Inertial Measurement Units (IMUs) attached to three different positions, namely the dominant wrist, the chest and the dominant sides ankle. Each IMU contains two 3-axis accelerometers, one 3-axis gyroscopes, one 3-axis magnetometers and one thermometer with the sampling rate of 100 Hz.

MHEALTH [9]: The MHEALTH dataset is also devised to benchmark methods of human activities recognition based on multimodal wearable sensor data. Three IMUs were respectively placed on the participants' chest, right wrist, and left ankle to record the acceleration (ms^{-2}), angular velocity (deg/s) and the magnetic field (local) data while they were performing 12 activities. The IMU on the chest also collected 2-lead ECG data (mV) to monitor the electrical activity of the heart. All sensing models are recorded at the frequency of 50 Hz.

MARS The MARS (Multimodal Activity Recognition with Sensing) dataset was collected while 8 participants (6 males, 2 females) were doing 5 basic activities (sitting, standing, walking, ascending stairs and descending stairs). Three IMU sensors, Phidget Spatial 3/3/3, were attached to the dominant wrist, the waist, and the dominant side's ankle, respectively, to collect the acceleration (gravitational acceleration g), angular velocity ($°/s$), and magnetism (nT). All IMUs collected the data at the frequency of 70 Hz.

Similar to [2], the experiments conducted on the two public datasets perform background activity recognition task [8]. The activities are categorized into 6 classes: lying, sitting/standing, walking, running, cycling and other activities. To tackle the task and ensure the rigorousness, all experiments are performed by Leave-One-Subject-Out (LOSO) cross-validation which can also test the person independence during the evaluation.

Model Implementation: In our experiments, we fix the size of glimpse to be 5×5 and deploy a CNN with two convolution layer, one ReLU layer and one max pooling layer. The dimensionality of $g\rho_t^f, gl_t^f, g_t^f, \theta_a$ and M are fixed to 128, 128, 220, 100 and 20 respectively. The location network is defined by a dual-component Gaussian with a variance of 0.22. The proposed method has two recurrent networks. One is the attention based LSTM with the cell size of 100. The number of time steps is 40, which defines the number of glimpses. The other one is the frame-based recurrent network with LSTM size of 1000 and time steps set to 5.

The proposed model is trained on a Nvidia Titan X Pascal GPU from scratch. The stochastic gradient descent with Adam update rule is used to minimize the cross-entropy loss function. The network parameters are optimized with a linearly annealed learning rate from 0.01 to 10^{-5}. To ensure the rigorousness, the experiments are performed by Leave-One-Subject-Out (LOSO) cross-validation.

B. Accuracy Comparison and Performance Analysis

To evaluate the performance of the proposed approach, our model, we conduct extensive experiments to compare its performance with the state-of-the-art methods on PAMAP2 and MHEALTH. We elaborately select other four state-of-the-art and multimodal feature-based approaches (MARCEL [2], FEM [11], CEM [22] and MKL [23]) and five baseline methods (Support Vector Machine (SVM), Random Forest(RF), K-Nearest Neighbors(KNN), Decision Tree(DT) and Single Neural Networks) to show the competitive power of the proposed method. To ensure fair comparison, the best parameters test, our model, is used on both datasets; the best trade-off parameter ($\lambda = 0.7$) is deployed for MARCEL; time-domain features including mean, variance, standard deviation, median and frequency-domain features including entropy and spectral entropy are utilized for FEM; each modality feature group are defined an independent kernel for MKL; and for other baseline methods, all modality features are deployed. All parameters adopted are in reference to the parameters suggested in literature. The results in Table I show the proposed our model outperforms all the state-of-the-art methods and the baseline methods.

TABLE I
COMPARISON AMONG OUR MODEL AND FOUR STATE-OF-THE-ART
METHODS AND FIVE BASELINE METHODS. FOR PAMAP2 DATASET,
ACCELEROMETER, GYROSCOPE AND MAGNETISM ARE UTILIZED. FOR
MHEALTH DATASET, ECG DATA ARE CONSIDERED ADDITIONALLY.

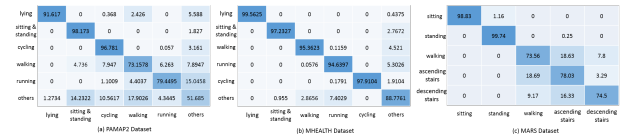| Method	Accuracy	
	PAMAP2	MHEALTH
MARCEL [2]	82.8	92.3
FEM+SVM [11]	76.4	70.7
CEM [22]	81	74.8
FEM+MKL [11], [23]	81.6	90.6
SVM	59.3	68.7
Random Forest	64.7	82.5
KNN	70.3	86.1
Decision Tree	57.8	78.7
Single Neural Networks	72.0	89.1
Our Model	**83.4**	**94.0**

Fig. 4. The confusion matrices of our model for background activity recognition on three datasets

To further explain the accuracy of our model on each specific activity, Figure 4 (a) and (b) show the confusion

matrices on both public datasets performing the background activity recognition task.

The results show the proposed approach performs well for most activities such as lying, sitting and standing, and cycling.

C. Effect of Activity Frames

We prove the effectiveness of our model on original features. To adapt features to the proposed model, multimodal features are stacked to form original frames, as Figure 2(a) shows. Table II presents feature extraction capability of activity frames, which shows that the proposed model based on the original frames outperforms most of the state-of-the-art methods (listed in Table I) even without activity frames. But utilizing the activity frames can improve the performance of original model due to the availability of the full relationship among features provided by activity frames.

TABLE II
FEATURE EXTRACTION CAPABILITY OF ACTIVITY FRAMES

	PAMAP2	MHEALTH	MARS
Original Frames	81.35	92.20	77.25
Activity Frames	**83.42**	**94.04**	**85.28**

D. Model Interpretability

One of the merits of our method is its interpretability. For wearable sensor-based activity recognition, subjects usually wear more than one sensors on their dominant body parts like arms, chest, and ankles, each sensor with multimodal. Attention mechanisms provide a superiority that it feeds the glimpse location back at each time step. Owing to the particularity of the activity frames, the attention model in our scenario not only provides the specific body parts it focuses on but also highlights the most contributing sensors and modals to diverse activities. In this section, we only present the experimental results of running, walking and lying down on MHEALTH dataset for simplicity. The available sensors on MHEALTH include ECG, chest accelerometer, ankle accelerometer, ankle gyroscope, ankle magnetometer, arm accelerometer, arm gyroscope and arm magnetometer. Figure 5 shows the glimpse heatmap for all sensors. Taking running as an example, we can observe ankle as the most active part of running. The chest also contributes a lot while arm involves the least. To further demonstrate the involvement of all sensory modal data, Table III concludes the percentage of our model "looking at" different modals for the latest 120 times (out of 200 times). It shows that for running, the most salient modal is ankle acceleration, which accounts for 30.55%. ECG and ankle angular velocity are also significant. The experimental results totally conform to the reality that while running, the most active body parts should be legs and ankles. Another self-evident truth is that in our experiments, one modality that can easily distinguish strenuous exercise like running from others is ECG. Also, since the model still "looks at" other modals for several times, it is able to better corroborate the claim that our model minimizes information loss.
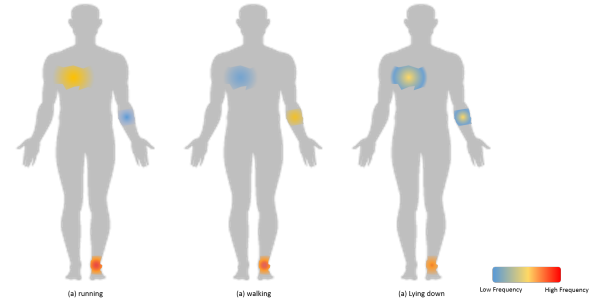

Fig. 5. Glimpse Heatmap
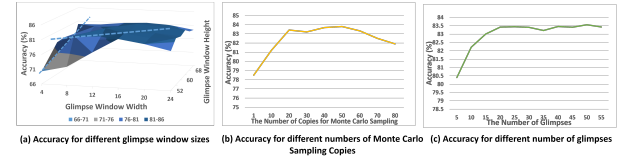
E. Hyper-Parameter Study

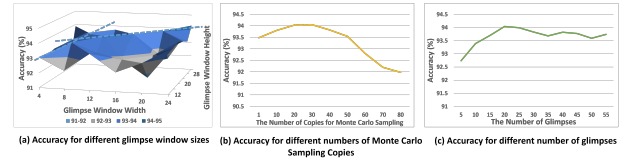Fig. 6. Experimental Results for Hyper-Parameter Tuning on PAMAP2

Fig. 7. Experimental Results for Hyper-Parameter Tuning on MHEALTH
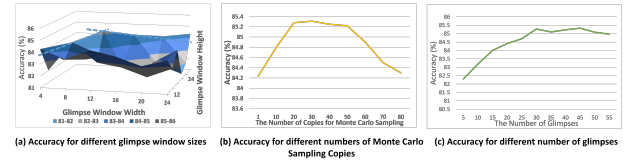

Fig. 8. Experimental Results for Hyper-Parameter Tuning on MARS

In this section, we mainly analyze three most contributing hyper-parameters to which the model is more sensitive in our experiments, namely the size of glimpse windows (width and height), the number of copies for Monte Carlo sampling and the number of glimpses. For the other hyper-parameters, we just use fixed empirical values as suggested in the previous subsection. The variation trend is shown as Figure 6, Figure 7 and Figure 8.

Taking Figure 6 as an example, firstly, we tune the width and the height of glimpse windows to figure out their relationship as shown in Figure 6 (a). Specifically, there are 13 3-axis vectors to present the temperature, accelerometer, gyroscope and magnetism in our experiments. After Algorithm 1, several 78x9 activity frames are generated. Figure 6 (a) shows that the accuracy achieves the best when the glimpse window size is 64x16 and there is an obvious "ridge" along which the whole figure is almost symmetric. All the points on the symmetric

TABLE III

MODALS INVOLVEMENTS ON MHEALTH DATASET (%). (ACC, GYRO, MAGN DENOTE ACCELERATION, ANGULAR VELOCITY AND MAGNETISM, RESPECTIVELY.)

activity	ECG	Acc_{chest}	Acc_{ankle}	$Gyro_{ankle}$	$Magn_{ankle}$	Acc_{arm}	$Gyro_{arm}$	$Magn_{arm}$
running	**21.98**	10.22	**30.55**	**15.86**	4.30	6.49	5.03	5.57
walking	7.23	11.05	**18.78**	**19.26**	8.72	**19.66**	9.46	5.83
lying down	6.58	13.45	**16.34**	10.23	**17.92**	10.29	10.72	**14.47**

line are in a ratio of $4 : 1$. This suggests that the approach favors a fixed ratio of the two dimensions of the glimpse window, in spite that we use the ratio of the activity frame size of $78 : 9$. Also, we can see that Figure 7 (a) and Figure 8 (a) both show the "ridge" while their optional glimpse window sizes are different because of different sizes of activity frames.

Figure 6 (b) and (c) show the experimental results of our studies on the effect of other two hyper-parameters, the number of copies for Monte Carlo sampling and the number of glimpses. In particular, Figure 6 (c) presents a trend that the accuracy increases remarkably at first and remains stable after getting a turning point. However, for the Monte Carlo sampling, too low or too high values lead to worse performance, as Figure 6 (b) shows. And we can notice that the variation trends in Figure 7 and Figure 8 enjoy the same patterns.

IV. CONCLUSION

This paper proposes an innovative HAR approach, our model, which includes (a) a novel form of multimodal sensor features, activity frames to fully extract relations between each pair of sensors and modality data and (b) an interpretable parallel recurrent model with convolutional attentions to combine recurrent attention and recurrent activity frames. The experiments show our method outperforms the state-of-the-art methods. Also, the method enjoys great interpretability in spite of the non-explanation of neural networks.

REFERENCES

[1] L. Yao, F. Nie, Q. Z. Sheng, T. Gu, X. Li, and S. Wang, "Learning from less for better: semi-supervised activity recognition via shared structure discovery," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 13–24.

[2] H. Guo, L. Chen, L. Peng, and G. Chen, "Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 1112–1123.

[3] D. Zhang, L. Yao, X. Zhang, W. Sen, C. Weitong, and R. Boots, "Cascade and parallel convolutional recurrent neural network," in *AAAI*, vol. 4, 2018, p. 12.

[4] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 33, 2014.

[5] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," in *Computer Vision, 1998. Sixth International Conference on*. IEEE, 1998, pp. 120–127.

[6] S. Y. Ooi, A. B. J. Teoh, Y. H. Pang, and B. Y. Hiew, "Image-based handwritten signature verification using hybrid methods of discrete radon transform, principal component analysis and probabilistic neural network," *Applied Soft Computing*, vol. 40, pp. 274–282, 2016.

[7] Z. Lai, Y. Xu, Q. Chen, J. Yang, and D. Zhang, "Multilinear sparse principal component analysis," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 10, pp. 1942–1950, 2014.

[8] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Wearable Computers (ISWC), 2012 16th International Symposium on*. IEEE, 2012, pp. 108–109.

[9] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J. A. Holgado-Terriza, S. Lee, H. Pomares, and I. Rojas, "Design, implementation and validation of a novel open framework for agile development of mobile health applications," *Biomedical engineering online*, vol. 14, no. 2, p. S6, 2015.

[10] K. Kunze and P. Lukowicz, "Dealing with sensor displacement in motion-based onbody activity recognition systems," in *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 2008, pp. 20–29.

[11] O. D. Lara, A. J. Pérez, M. A. Labrador, and J. D. Posada, "Centinela: A human activity recognition system based on acceleration and vital sign data," *Pervasive and mobile computing*, vol. 8, no. 5, pp. 717–729, 2012.

[12] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *IEEE Transactions on information technology in biomedicine*, vol. 10, no. 1, pp. 119–128, 2006.

[13] E. M. Tapia, S. S. Intille, W. Haskell, K. Larson, J. Wright, A. King, and R. Friedman, "Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor," in *Wearable Computers, 2007 11th IEEE International Symposium on*. IEEE, 2007, pp. 37–40.

[14] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1307–1310.

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

[16] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," *The Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[18] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas, "Learning where to attend with deep architectures for image tracking," *Neural computation*, vol. 24, no. 8, pp. 2151–2184, 2012.

[19] A. Haque, A. Alahi, and L. Fei-Fei, "Recurrent attention models for depth-based person identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1229–1238.

[20] M. Zontak, I. Mosseri, and M. Irani, "Separating signal from noise using patch recurrence across scales," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1195–1202.

[21] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[22] H. Guo, L. Chen, Y. Shen, and G. Chen, "Activity recognition exploiting classifier level fusion of acceleration and physiological signals," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 2014, pp. 63–66.

[23] S. Althloothi, M. H. Mahoor, X. Zhang, and R. M. Voyles, "Human activity recognition using multi-features and multiple kernel learning," *Pattern recognition*, vol. 47, no. 5, pp. 1800–1812, 2014.