Multi-agent Attentional Activity Recognition

Kaixuan Chen[1] , **Lina Yao**[1] , **Dalin Zhang**[1] , **Bin Guo**[2] and **Zhiwen Yu**[2]
[1]University of New South Wales
[2]Northwestern Polytechnical University
{kaixuan.chen@student., lina.yao@}unsw.edu.au,

Abstract

Multi-modality is an important feature of sensor based activity recognition. In this work, we consider two inherent characteristics of human activities, the spatially-temporally varying salience of features and the relations between activities and corresponding body part motions. Based on these, we propose a multi-agent spatial-temporal attention model. The spatial-temporal attention mechanism helps intelligently select informative modalities and their active periods. And the multiple agents in the proposed model represent activities with collective motions across body parts by independently selecting modalities associated with single motions. With a joint recognition goal, the agents share gained information and coordinate their selection policies to learn the optimal recognition model. The experimental results on four real-world datasets demonstrate that the proposed model outperforms the state-of-the-art methods.

1 Introduction

The ability to identify human activities via on-body sensors has been of interest to the healthcare community [Anguita *et al.*, 2013], the entertainment [Freedman and Zilberstein, 2018] and fitness [Guo *et al.*, 2017] community. Some works of Human Activity Recognition (HAR) are based on handcrafted features for statistical machine learning models [Lara and Labrador, 2013]. Until recently, deep learning has experienced massive success in modeling high-level abstractions from complex data [Pouyanfar *et al.*, 2018], and there is a growing interest in developing deep learning for HAR [Hammerla *et al.*, 2016]. Despite this, these methods still lack sufficient justification when being applied to HAR. In this work, we consider two inherent characteristics of human activities and exploit them to improve the recognition performance.

The first characteristic of human activities is the spatially-temporally varying salience of features. Human activities can be represented as a sequence of multi-modal sensory data. The modalities include acceleration, angular velocity and magnetism from different positions of testers' bodies, such as chests, arms and ankles. However, only a part of modalities from specific positions are informative for recognizing certain activities [Wang and Wang, 2017]. Irrelevant modalities often influence the recognition and undermine the performance. For instance, identifying lying mainly relies on people's orientations (magnetism), and going upstairs can be easily distinguished by upward acceleration from arms and ankles. In addition, the significance of modalities changes over time. Intuitively, the modalities are only important when the body parts are actively participating in the activities. Therefore, we propose a spatial-temporal attention method to select salient modalities and their active periods that are indicative of the true activity. Attention has been proposed as a sequential decision task in earlier works [Denil *et al.*, 2012; Mnih *et al.*, 2014]. This mechanism has been applied to sensor based HAR in recent years. [Chen *et al.*, 2018] and [Zhang *et al.*, 2018] transform the sensory sequences into 3-D activity data by replicating and permuting the input data, and they propose to attentionally keep a focal zone for classification. However, these methods heavily rely on data preprocessing, and the replication increases the computation complexity unnecessarily. Also, these methods do not take the temporally-varying salience of modalities into account. In contrast, the proposed spatial-temporal attention approach directly selects informative modalities and their active time that are relevant to classification from raw data. The experiment results shows that our model makes HAR more explainable.

The second characteristic of human activities considered in this paper is activities are portrayed by motions on several body parts collectively. For instance, running can be seen as a combination of arm and ankle motions. Some works like [Radu *et al.*, 2018; Yang *et al.*, 2015] are committed to fusing multi-modal sensory data for time-series HAR, but they only fuse the information of local modalities from the same positions. These methods, as well as the existing attention based methods [Chen *et al.*, 2018; Zhang *et al.*, 2018], are limited in capturing the global interconnections across different body parts. To fill this gap, we propose a multi-agent reinforcement learning approach. We simplify activity recognition by dividing the activities into sub-motions with which an independent intelligent agent is associated and by coordinating the agents' actions. These agents select informative modalities independently based on both their local observations and the information shared by each other. Each agent can individually learn an efficient selection policy by trial-and-error. After a sequence of selec-

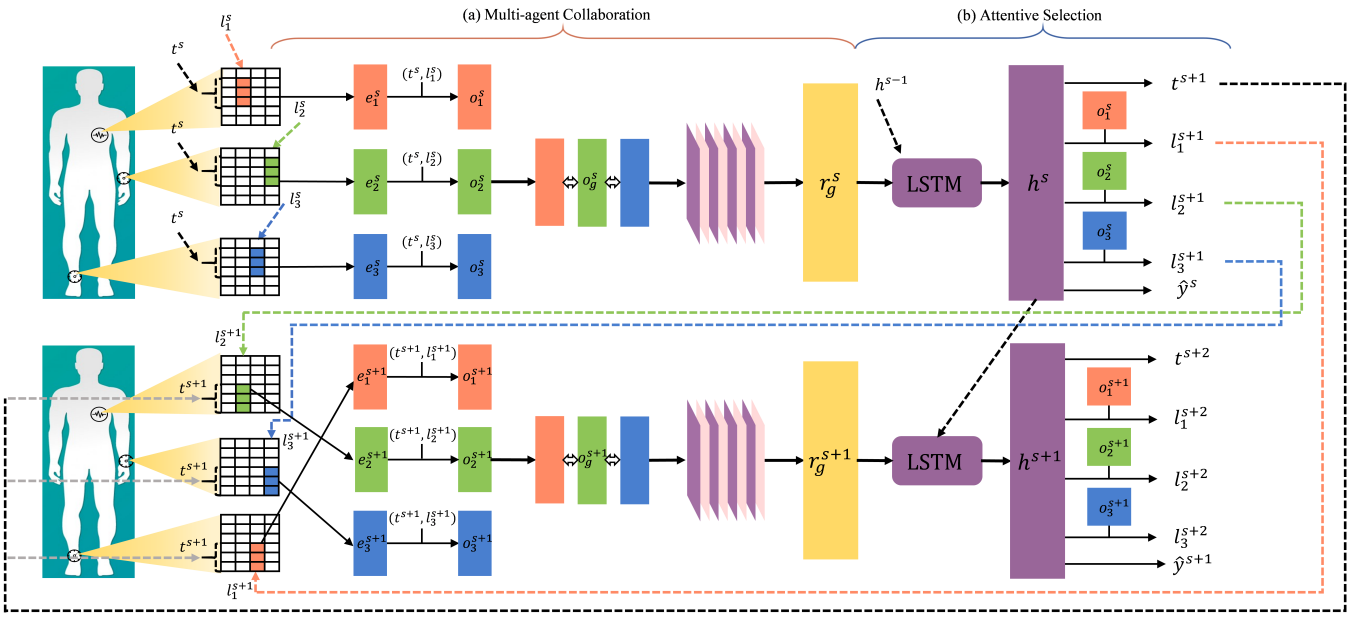


Figure 1: The overview of the proposed model. At each step s, three agents a_1, a_2, a_3 individually select modalities and obtain observations o_1^s, o_2^s, o_3^s from the input $\mathbf{x}$ at (t^s, l_1^s), (t^s, l_2^s) and (t^s, l_3^s). The agents then exchange and process the gained information to get the representation r_g^s of the shared observation. And they decide the next locations again. Based on a sequence of observations after an episode, the agents jointly make the classification. Red, green and blue denote the workflows that are associated with a_1, a_2, a_3, respectively. Other colors denote the shared information and its representations.

tions and information exchanges, a joint decision on recognition is made. The selection policies are incrementally coordinated during training since the agents share a common goal which is to minimize the loss caused by false recognition.

The key contributions of this research are as follows:

- We propose a spatial-temporal attention method for temporal sensory data, which considers the spatially-temporally varying salience of features, and allows the model to focus on the informative modalities that are only collected in their active periods.
- We propose a multi-agent collaboration method. The agents represent activities with collective motions by independently selecting modalities associated with single motions and sharing observations. The whole model can be optimized by coordinating the agents' selection policies with the joint recognition goal.
- We evaluate the proposed model on four datasets. The comprehensive experiment results demonstrate the superiority of our model to the state-of-the-art approaches.

2 The Proposed Method

2.1 Problem Statement

We now detail the human activity recognition problem on multi-modal sensory data. Each input sample $(\mathbf{x}, y)$ consists of a 2-d vector $\mathbf{x}$ and an activity label y. Let $\mathbf{x} = [\mathbf{x}^0, \mathbf{x}^1, ...\mathbf{x}^K]$ where K denotes the time window length and $\mathbf{x}^i$ denotes the sensory vector collected at the point i in time. $\mathbf{x}^i$ is the combination of multi-modal sensory data collected from testers' different body positions such as chests, arms and ankles. Suppose that $\mathbf{x}^i = (\mathbf{m}_1^i, \mathbf{m}_2^i, ...\mathbf{m}_N^i) = (x_0^i, x_1^i, ...x_P^i)$, where $\mathbf{m}$ denotes data collected from each position, N denotes the number of positions (in our datasets, $N = 3$), and P denotes the number of values per vector. Therefore, $\mathbf{x} \in R^{K \times P}$ and $y \in [1, ..., C]$. C represents the number of activity classes. The goal of the proposed model is to predict the activity y.

2.2 Model Structure

The overview of the model structure is shown in Figure 1. At each step s, the agents select an active period together and individually select informative modalities from the input $\mathbf{x}$. These agents share their information and independently decide where to "look at" at the next step. The locations are determined spatially and temporally in terms of modalities and time. After several steps, the final classification is jointly conducted by the agents based on a sequence of the observations. Each agent can incrementally learn an efficient decision policy over episodes. But by having the same goal, which is to jointly minimize the recognition loss, they collaborate with each other and learn to align their behaviors such that it achieves their common goal.

Multi-agent Collaboration

In this work we simplify activity recognition by dividing the activities into sub-motions and require each agent select informative modalities that are associated with one motion. Suppose that we employ H agents $a_1, a_2, ...a_H$ (we assume $H = 3$ in this paper for simplicity). The workflows of a_1, a_2, a_3 are shown in red, green and blue in Figure 1.

At each step s, each agent locally observes a small patch of

x, which includes information of a specific modality from a motion in its active period. Let the observations be e_1^s, e_2^s, e_3^s as Figure 1 shows. They are extracted from **x** at the locations (t^s, l_1^s), (t^s, l_2^s) and (t^s, l_3^s), respectively, where t denotes the selected active period and l denote the location of a modality in the input **x**. The model encodes the region around (t^s, l_i^s) ($i \in \{1, 2, 3\}$) with high resolution but uses a progressively lower resolution for points further from (t^s, l_i^s) in order to remove noises and avoid information loss in [Zontak *et al.*, 2013]. We then further encode the observations into higher level representations. With regard to each agent a_i ($i \in \{1, 2, 3\}$), the observation e_i^s and the location (t^s, l_i^s) are linear transformed independently, parameterized by θ_e and θ_{tl}, respectively. Next, the summation of these two parts is further transformed with another linear layer parameterized by θ_o The whole process can be summarized as the following equation:

$$\begin{aligned} o_i^s &= f_o(e_i^s, t^s, l_i^s; \theta_e, \theta_{tl}, \theta_o) \\ &= L(L(e_i^s) + L(concat(t^s, l_i^s))) \;\; i \in \{1, 2, 3\}, \end{aligned} \tag{1}$$

where $L(\bullet)$ denotes a linear transformation and $concat(t^s, l_i^s)$ represents the concatenation of t^s and l_i^s. Each linear layer is followed by a rectified linear unit (ReLU) activation. Therefore, o_i^s contains information from "what" ($\rho(C^f, l_t^f)$), "where" (l_t^f) and "when".

Making multiple observations not only avoids the system processing the whole data at a time but also maximally prevents the information loss from only selecting one region of data. Furthermore, multiple agents make observations individually so that they can represent activities with the collective modalities from different motions. The model can explore various combinations of modalities to recognize activities during learning.

Then we are interested in the collaborative setting where the agents communicate with each other and share the observations they make. So we get the shared observation o_g^s by concatenate o_1^s, o_2^s, o_3^s together.

$$o_g^s = concat(o_1^s, o_2^s, o_3^s), \tag{2}$$

so that o_g^s contains all the information observed by three agents. A convolutional network is further applied to process o_g^s and extract the informative spatial relations. The output is then reshaped to be the representation r_g^s.

$$r_g^s = f_c(o_g^s; \theta_c) = reshape(Conv(o_g^s)) \tag{3}$$

And r_g^s represents the activity to be identified with multiple modalities selected from motions on different body positions.

Attentive Selection

In this section, the details about how to select modalities and active period attentively are introduced. We first introduce the episodes in this work. The agents incrementally learn the attentive selection policies over episodes. In each episode, following the bottom-up processes, the model attentively selects data regions and integrates the observations over time to generate dynamic representations, in order to determine effective selections and maximize the rewards, i.e., minimize the loss. Based on this, LSTM is appropriate to build an episode as it incrementally combines information from time steps to obtain final results. As can be seen in Figure 1, at each step s, the LSTM module receives the representation r_g^s and the previous hidden state h^{s-1} as the inputs. Parameterized by θ_h, it outputs the current hidden state h^s:

$$h^s = f_h(r_g^s, h^{s-1}; \theta_h) \tag{4}$$

Now we introduce the selection module. The agents are supposed to select salient modalities and an active period at each step. To be specific, they need to select the locations where they make next observations. Three agents control $l_1^{s+1}, l_1^{s+1}, l_3^{s+1}$ independently based on both the hidden state h^s and their individual observations o_1^s, o_2^s, o_3^s so that the individual decisions are made from the overall observation as well. t^{s+1} is jointly decided based on h^s only since it is a common selection. The decisions are made by the agents' selection policies which are defined by Gaussian distribution stochastic process:

$$l_i^{s+1} \sim P(\cdot \mid f_l(h^s, o_i^s; \theta_{l_i})) \;\; i \in \{1, 2, 3\}, \tag{5}$$

and

$$t^{s+1} \sim P(\cdot \mid f_t(h^s; \theta_t)) \tag{6}$$

The purpose of stochastic selections is to explore more kinds of selection combinations such that the model can learn the best selections during training.

To align the agents' selection policies, we assign the agents a common goal that correctly recognizing activities after a sequence of observations and selections. They together receive a positive reward if the recognition is correct. Therefore, at each step s, a prediction $\hat{y}^s$ is made by:

$$\hat{y}^s = f_y(h^s; \theta_y) = softmax(L(h^s)) \tag{7}$$

Usually, agents receive a reward r after each step. But in our case, since only the classification in the last step S is representative, the agents receive a delayed reward R after each episode.

$$R = \begin{cases} 1 \;\; if \;\; \hat{y}^S = y \\ 0 \;\; if \;\; \hat{y}^S \neq y \end{cases} \tag{8}$$

The target of optimization is to coordinate all the selection policies by maximizing the expected value of the reward $\bar{R}$ after several episodes.

2.3 Training and Optimization

This model involves parameters that define the multi-agent collaboration and the attentive selection. The parameters $\Theta = \{\theta_e, \theta_{tl}, \theta_o, \theta_c, \theta_h, \theta_{l_i}, \theta_t, \theta_y\}$ ($i \in \{1, 2, 3\}$). The parameters for classification can be optimized by minimizing the cross-entropy loss:

$$L_c = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_n(c) \log F_y(\mathbf{x}), \tag{9}$$

where F_y is the overall function that outputs $\hat{y}$ given **x**. C is the number of activity classes, and $y_n(c) = 1$ if the n-th sample belongs to the c-th class and 0 otherwise.

However, selection policies that are mainly defined by θ_{l_i} ($i \in \{1, 2, 3\}$) and θ_t are expected to select a sequence of

locations. The parameters are thus non-differentiable. In this view, we deploy a Partially Observable Markov Decision Process (POMDP) [Cai *et al.*, 2009] to solve the optimization problem. Suppose $e^s = (e_1^s, e_2^s, e_3^s)$, $lt^s = (l_1^s, l_2^s, l_3^s, t^s)$, We consider each episode as a trajectory $\tau = \{e^1, lt^1, y^1; e^2, lt^2, y^2; ..., e^S, lt^S, y^S\}$. Each trajectory represents one order of the observations, the locations and the predictions our agents make. After agents repeat N episodes, we can obtain $\{\tau^1, \tau^2, ..., \tau^N\}$, and each τ has a probability $p(\tau; \Theta)$ to be obtained. The probability depends on the selection policy $\Pi = (\pi_1, \pi_2, \pi_3)$ of the agents.

Our goal is to learn the best selection policy Π that maximizes $\bar{R}$. Specifically, Π is decided by Θ. Thus we need to find out the optimized $\Theta^* = \arg\max_{\Theta}[\bar{R}]$. One common way is gradient ascent.

Generally, given a sample x with reward $f(x)$ and probability $p(x)$, the gradient can be calculated as follows:

$$\begin{aligned}\nabla_\theta E_x[f(x)] &= \nabla_\theta \sum_x p(x) f(x) \\ &= \sum_x p(x) \frac{\nabla_\theta p(x)}{p(x)} f(x) \\ &= \sum_x p(x) \nabla_\theta log p(x) f(x) \\ &= E_x[f(x) \nabla_\theta log p(x)] \end{aligned} \tag{10}$$

In our case, a trajectory τ can be seen as a sample, the probability of each sample is $p(\tau; \Theta)$, and the reward function $\bar{R} = E_{p(\tau;\Theta)}[R]$. We have the gradient:

$$\nabla_\Theta \bar{R} = E_{p(\tau;\Theta)}[R \nabla_\Theta log p(\tau; \Theta)] \tag{11}$$

By considering the training problem as a POMDP and following the REINFORCE rule [Williams, 1992]:

$$\nabla_\Theta \bar{R} = E_{p(\tau;\Theta)}[R \sum_{s=1}^{S} \nabla_\Theta log \Pi(y|\tau_{1:s}; \Theta)] \tag{12}$$

Since we need several samples τ for one input **x** to learn the best policy combination, we adopt Monte Carlo sampling which utilizes randomness to yield results that might be theoretically deterministic. Supposing M is the number of Monte Carlo sampling copies, we duplicate the same input for M times and average the prediction results. The M copies generate M subtly different results owing to the stochasticity, so we have:

$$\nabla_\Theta \bar{R} \approx \frac{1}{M} \sum_{i=1}^{M} R^{(i)} \sum_{s=1}^{S} \nabla_\Theta log \Pi(y^{(i)}|\tau_{1:s}^{(i)}; \Theta), \tag{13}$$

where M denotes the number of Monte Carlo samples, i denotes the i^{th} duplicated sample, and y_i is the correct label for the i^{th} sample. Therefore, the overall optimization can be summarized as maximizing $\bar{R}$ and minimizing Eq. 9. The detailed procedure is shown in Algorithm 1.

Algorithm 1 Training and Optimization

Require: sensory matrix **x**, label y,
the length of episodes S,
the number of Monte Carlo samples M.
Ensure: parameters Θ.

$\Theta = RandomInitialize()$
while training **do**
  duplicate **x** for M times
  for i from 1 to M **do**
    $l_1^{1(i)}, l_2^{1(i)}, l_3^{1(i)}, t^{1(i)} = RandomInitialize()$
    for s from 1 to S **do**
      extract $e_1^{s(i)}, e_2^{s(i)}, e_2^{s(i)}$
      $o_1^{s(i)}, o_2^{s(i)}, o_3^{s(i)} \leftarrow Eq.\ 1$
      $o_g^{s(i)}, r_g^{s(i)}, h^{s(i)} \leftarrow Eq.\ 2, Eq.\ 3, Eq.\ 4$
      $l_1^{s(i)}, l_2^{s(i)}, l_3^{s(i)}, t^{s(i)} \leftarrow Eq.\ 5, Eq.\ 6$
      $\hat{y}^{s(i)} \leftarrow Eq.\ 7$
      record $\tau_{1:s}^{(i)}$
    end for
    $R^{(i)} \leftarrow Eq.\ 8$
  end for
  $\hat{y} = \frac{1}{M} \sum_{i=1}^{M} \hat{y}^{S(i)}$
  $L_c, \nabla_\Theta \bar{R} \leftarrow Eq.\ 9, Eq.\ 13$
  $\Theta \leftarrow \Theta - \nabla_\Theta L_c + \nabla_\Theta \bar{R}$
end while
return Θ

3 Experiments

3.1 Experiment Setting

We now introduce the settings in our experiments. The time window of inputs is 20 with 50% overlap. The size of each observation patch is set to $\frac{K}{8} \times \frac{P}{8}$, where $K \times P$ is the size of the inputs. In the partial observation part, the sizes of $\theta_e, \theta_{tl}, \theta_o$ are $128, 128, 220$, respectively. The filter size of the convolutional layer in the shared observation module is $1 \times M$ and the number of feature maps is 40, where M denotes the width of o_g^s. The size of LSTM cells is 220, and the length of episodes is 40. The Gaussian distribution that defines the selection policies is with a variance of 0.22.

To ensure the rigorousness, the experiments are performed by Leave-One-Subject-Out (LOSO) on four datasets, MHEALTH [Banos *et al.*, 2014], PAMAP2 [Reiss and Stricker, 2012], UCI HAR [Anguita *et al.*, 2013] and MARS. They contain $10, 9, 30, 8$ subjects' data, respectively.

3.2 Comparison with State-of-the-Art

To verify the overall performance of the proposed model, we first compare our model with other state-of-the-art methods. The compared methods include a convolutional model on multichannel time series for HAR (MC-CNN) [Yang *et al.*, 2015], a CNN-based multi-modal fusion model (C-Fusion) [Radu *et al.*, 2018], a deep multimodal HAR model with classifier ensemble (MARCEL) [Guo *et al.*, 2016], an ensemble of deep LSTM learners for activity recognition (E-LSTM) [Guan and Plötz, 2017], a parallel recurrent model with convolutional attentions (PRCA) [Chen *et al.*, 2018] and a weighted average spatial LSTM with selective attention (WAS-LSTM) [Zhang *et al.*, 2018].

As can be observed in Table 1, with respect to the datasets, MARCEL, E-LSTM, PRCA, WAS-LSTM and the proposed model perform better than MC-CNN and C-Fusion

	Method	MC-CNN	C-Fusion	MARCEL	E-LSTM	PRCA*	WAS-LSTM*	**Ours***
MH	Accuracy	87.19±0.77	88.66±0.62	92.35±0.46	91.58±0.38	93.32±0.75	91.42±1.25	**96.12±0.37**
	Precision	86.50±0.61	86.36±0.72	93.17±0.84	90.50±0.68	92.11±0.96	91.35±0.70	**95.46±0.33**
	Recall	87.29±0.44	89.68±0.72	92.81±0.44	91.58±0.59	92.25±0.94	91.99±1.04	**96.76±0.30**
	F1	86.89±0.66	87.98±0.79	92.98±0.74	91.03±0.68	92.17±1.06	91.66±1.21	**96.10±0.47**
	Method	MC-CNN	C-Fusion	MARCEL	E-LSTM	PRCA*	WAS-LSTM*	**Ours***
PMP	Accuracy	81.16±1.32	81.86±0.74	82.87±0.81	83.21±0.68	82.39±1.04	84.89±2.18	**90.33±0.62**
	Precision	81.57±0.89	81.63±0.53	83.51±0.71	84.01±0.54	82.44±0.99	84.44±1.54	**89.25±0.78**
	Recall	81.43±0.64	81.96±0.89	81.12±0.79	83.88±0.74	82.86±0.90	84.20±1.83	**90.49±0.94**
	F1	81.50±0.72	81.79±0.71	82.29±0.76	83.94±0.95	82.64±1.19	84.81±1.06	**89.86±0.81**
	Method	MC-CNN	C-Fusion	MARCEL	E-LSTM	PRCA*	WAS-LSTM*	**Ours***
HAR	Accuracy	75.86±0.59	74.64±0.78	80.16±0.72	80.78±0.94	81.29±1.22	71.29±1.08	**85.72±0.83**
	Precision	76.93±0.78	73.30±0.75	81.63±0.50	81.34±0.43	80.55±1.26	70.76±0.93	**85.61+0.53**
	Recall	75.81±0.39	74.07±0.48	80.81±0.64	80.63±0.54	81.66±1.03	71.10±1.37	**85.08±0.72**
	F1	76.36±1.11	73.68±0.79	81.21±0.85	80.98±0.64	81.11±1.02	70.92±1.16	**85.34±0.58**
	Method	MC-CNN	C-Fusion	MARCEL	E-LSTM	PRCA*	WAS-LSTM*	**Ours***
MARS	Accuracy	81.34±0.59	81.48±0.56	81.68±0.87	81.59±0.77	85.38±0.82	74.82±1.42	**88.29±0.87**
	Precision	81.68±0.62	81.84±0.68	81.23±0.84	81.79±0.85	85.99±1.07	75.89±1.54	**88.75±0.81**
	Recall	81.06±0.90	82.15±0.82	82.44±0.54	81.65±0.93	84.95±0.95	74.80±1.63	**87.20±0.67**
	F1	81.32±0.42	81.99±0.64	81.85±0.97	81.71±0.81	85.46±1.02	75.34±1.27	**87.96±0.74**

Table 1: The prediction performance of the proposed approach and other state-of-the-art methods. * denotes attention based state-of-the-art. The best performance is indicated in bold.

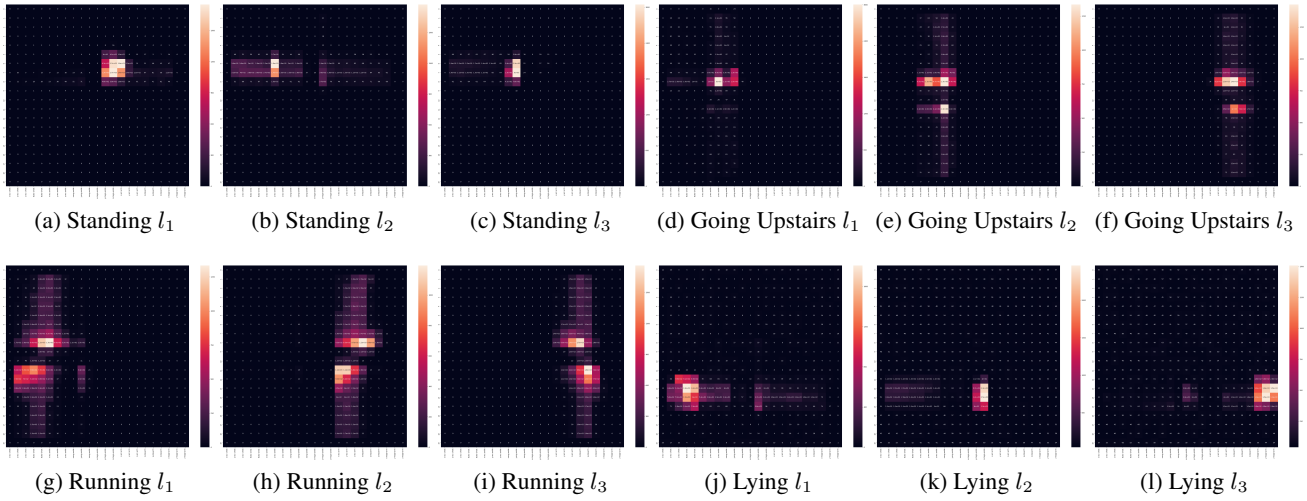
Ablation	MHEALTH				PAMAP2			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
S1	85.75±0.70	84.67±0.92	85.83±0.78	84.34±0.89	79.60±0.56	79.86±0.51	79.68±0.49	79.57±0.83
S2	80.59±1.54	80.37±0.95	80.95±1.44	80.00±1.12	71.49±1.55	71.62±1.18	71.38±1.36	71.49±1.42
S3	85.49±0.88	85.67±0.35	84.62±0.71	85.14±0.86	77.68±0.73	77.35±0.52	77.74±0.82	77.04±0.59
S4	88.32±0.75	87.11±0.96	87.25±0.94	87.17±1.06	78.39±1.04	78.44±0.99	78.86±0.90	78.64±1.19
S5	91.93±0.94	90.85±0.85	91.35±0.73	91.88±0.81	83.53±0.95	83.66±0.85	83.38±0.61	83.51±0.74
S6	**96.12±0.37**	**95.46±0.33**	**96.76±0.30**	**96.10±0.47**	**90.33±0.62**	**89.25±0.78**	**90.49±0.94**	**89.86±0.81**
Ablation	UCI HAR				MARS			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
S1	73.68±0.73	73.79±0.54	73.98±0.59	73.38±0.42	79.78±0.89	79.37±0.82	79.49±0.42	79.31±0.64
S2	68.95±2.85	67.41±2.97	67.92±2.88	67.66±2.81	71.94±2.18	71.34±2.52	71.99±2.67	71.15±2.71
S3	75.45±0.77	75.54±1.27	75.49±0.88	75.88±0.94	78.34±0.84	78.42±0.83	78.48±0.98	78.40±0.99
S4	76.29±1.22	76.55±1.26	77.66±1.03	77.11±1.02	81.38±0.82	81.99±1.07	81.95±0.95	81.46±1.02
S5	80.71±0.94	80.95±1.82	80.44±0.74	80.11±0.52	85.51±0.86	84.94±0.73	85.73±0.66	84.81±0.98
S6	**85.72±0.83**	**85.61+0.53**	**85.08±0.72**	**85.34±0.58**	**88.29±0.87**	**88.75±0.81**	**87.20±0.67**	**87.96±0.74**

Table 2: Ablation Study. S1 $\sim$ S6 are six structures by systematically removing five components from the proposed model. The considered components are: a) the selection module, (b) the partial observation processing from e_1^s to o_i^s ($i \in \{1, 2, 3\}$), (c) the convolutional merge of shared observations, (d) the temporal attentive selection (e) the multi-agent for selection.

in MHEALTH and PAMAP2, as these models enjoy higher variance. They fit well when data contain numerous features and complex patterns. On the other hand, data in UCI HAR and MARS have fewer features, but MARCEL, E-LSTM, PRCA and our model still perform well while the performance of WAS-LSTM deteriorates. The reason is that WAS-LSTM is based on a complex structure and it requires more features as input. In contrast, MARCEL and E-LSTM adopt rather simple models like DNNs and LSTMs. Despite the ensembles, they are still suitable for fewer features. PRCA and the proposed model select salient features directly with intuitive rewards, so they do not necessarily need a large number of features as well. In addition, the attention based methods, PRCA and WAS-LSTM, are more unstable than the other methods since the selection is stochastic and they cannot guarantee the effectiveness of all the selected features. Overall, our model outperforms the compared state-of-the-art and eliminates the instability of regular selective attentions.

3.3 Ablation Study

We perform a detailed ablation study to examine the contributions of the proposed model components to the prediction performance in Table 2. Considering that there are five removable components in this model: (a) the modality selection module, (b) the transformation from e_i^s to o_i^s ($i \in \{1, 2, 3\}$), (c) the convolutional network for higher-level representations, (d) the temporal attentive selection (e) the multi-agent. We consider six structures: **S1**: We first remove the selection module including the observations, episodes, selections and rewards. For comparison, we set S1 to be a regular CNN as



(a) Standing l_1 (b) Standing l_2 (c) Standing l_3 (d) Going Upstairs l_1 (e) Going Upstairs l_2 (f) Going Upstairs l_3

(g) Running l_1 (h) Running l_2 (i) Running l_3 (j) Lying l_1 (k) Lying l_2 (l) Lying l_3

Figure 2: Visualization of the selected modalities and time on MHEALTH. The input matrices' size is 20×23, where 20 is the length of the time window and 23 is the number of modalities. Thus each grid denotes an input feature, and the values in the grids represent the frequency with which this feature is selected. Lighter colors denote higher frequency. To be clear, detailed illustration is provided in Table 3.

a baseline. **S2**: We employ one agent but remove (b), (c), (d) and (e), so the workflow is: inputs $\rightarrow e_i^s \rightarrow$ LSTM $\rightarrow$ selections and rewards. The performance decreases and is more unstable than other structures. Although S2 includes attentions, the model does not include the previous selections in their observations, which influences their next decisions significantly. **S3**: Based on S2, we add (b) to (a). (b) contributes considerably since the prediction results are improved by 5% to 7%, because it feeds back the history selections to the agents for learning. **S4**: We further consider (c) in the model. It can be observed that this setting also achieves better performance than S3 since it convolutionally merges the partial observations. **S5**: (d) is added. The workflow is the same as S3, but the agents make an additional action: selecting t^S, which leads to another attention mechanism in time level. The performance is improved by 3% to 5%. **S6**: The proposed model. When combining all these benefits, our model achieves the best performance, higher than S5 by 5% to 7%.

3.4 Visualization and Explainability

The proposed method decomposes the activities into participating motions, from each of which the agents decide the most salient modalities individually, which makes the model explainable. We present the visualized process of recognizing standing, going upstairs, running and lying on MHEALTH. The available features include three 3-axis acceleration from chests, arms and ankles, two ECG signals, two 3-axis angular velocity and two 3-axis magnetism vectors from arms and ankles. Figure 2 shows the modality heatmaps of all agents. We observe that each agent does focus on only a part of modalities in a time period during recognition. Table 3 lists the most frequently selected modalities. We can observe that magnetism (orientation) in standing and lying is selected as one of the most active features, owing to the fact it is easy to distinguish between standing and lying with people's orientation. Another example is that the most distinguishing characteristic of going upstairs is "up". Therefore, Z-axis acceleration is specifically selected by agents for going upstairs. Also, identifying running involves acceleration, ECG, and arm swing, which conforms to the experiment evidence as well. The agents also select several other features with lower frequencies, which avoids losing effective information.

Activity	Agent	Location
Standing	1	Y,**Z**-Magn-Ankle, X,Y-Acc-Arm
	2	**Y**-Acc-Ankle, ECG1,2-Chest
	3	X,**Y**-Ang-Ankle
Going Upstairs	1	Y,**Z**-Acc-Ankle, X,Y-Ang-Ankle
	2	X,Y,Z-Acc-Ankle, **X**,Y-Ang-Ankle
	3	Y,**Z**-Acc-Arm, X,Y,Z-Ang-Arm
Running	1	Y,Z-Acc-Chest, ECG1,2-Chest **X**,Y,Z-Acc-Ankle
	2	X,Y,Z-Acc-Arm, **X**,Y-Ang-Arm
	3	Y,Z-Acc-Arm, X,**Y**.Z-Ang-Arm
Lying	1	Y,Z-Acc-Chest, **ECG1**,2-Chest
	2	Y,**Z**-Magn-Ankle
	3	X,**Y**,Z-Magn-Arm

Table 3: The active modalities for activities selected by the agents are listed. X, Y, Z denote the axis of data. Acc, Ang and Magn denote acceleration, angular velocity and magnetism, respectively. The most frequently selected locations are indicated in bold.

4 Conclusion

In this work, we first propose a selective attention method for spatially-temporally varying salience of features. Then, multi-agent is proposed to represent activities with collective motions. The agents' cooperate by aligning their actions to achieve their common recognition target. We experimentally evaluate our model on four real-world datasets, and the results validate the contributions of the proposed model.

References

[Anguita *et al.*, 2013] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *ESANN*, 2013.

[Banos *et al.*, 2014] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid: a novel framework for agile development of mobile health applications. In *International Workshop on Ambient Assisted Living*, pages 91–98. Springer, 2014.

[Cai *et al.*, 2009] Chenghui Cai, Xuejun Liao, and Lawrence Carin. Learning to explore and exploit in pomdps. In *Advances in Neural Information Processing Systems (NIPS)*, pages 198–206, 2009.

[Chen *et al.*, 2018] Kaixuan Chen, Lina Yao, Xianzhi Wang, Dalin Zhang, Tao Gu, Zhiwen Yu, and Zheng Yang. Interpretable parallel recurrent neural networks with convolutional attentions for multi-modality activity modeling. In *Neural Networks (IJCNN), 2018 International Joint Conference on*, pages 3016–3021. IEEE, 2018.

[Denil *et al.*, 2012] Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. Learning where to attend with deep architectures for image tracking. *Neural computation*, 24(8):2151–2184, 2012.

[Freedman and Zilberstein, 2018] Richard Gabriel Freedman and Shlomo Zilberstein. Roles that plan, activity, and intent recognition with planning can play in games. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Guan and Plötz, 2017] Yu Guan and Thomas Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, IMWUT*, 1(2):11, 2017.

[Guo *et al.*, 2016] Haodong Guo, Ling Chen, Liangying Peng, and Gencai 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, UbiComp 2016, Heidelberg, Germany, September 12-16, 2016*, pages 1112–1123, 2016.

[Guo *et al.*, 2017] Xiaonan Guo, Jian Liu, and Yingying Chen. Fitcoach: Virtual fitness coach empowered by wearable mobile devices. In *2017 IEEE Conference on Computer Communications, INFOCOM 2017, Atlanta, GA, USA, May 1-4, 2017*, pages 1–9, 2017.

[Hammerla *et al.*, 2016] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 1533–1540, 2016.

[Lara and Labrador, 2013] Oscar D Lara and Miguel A Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials*, 15(3):1192–1209, 2013.

[Mnih *et al.*, 2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[Pouyanfar *et al.*, 2018] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and SS Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):92, 2018.

[Radu *et al.*, 2018] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, IMWUT*, 1(4):157, 2018.

[Reiss and Stricker, 2012] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 108–109. IEEE, 2012.

[Wang and Wang, 2017] Hongsong Wang and Liang 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.

[Williams, 1992] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[Yang *et al.*, 2015] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *IJCAI*, pages 3995–4001, 2015.

[Zhang *et al.*, 2018] Xiang Zhang, Lina Yao, Chaoran Huang, Sen Wang, Mingkui Tan, Guodong Long, and Can Wang. Multi-modality sensor data classification with selective attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 3111–3117, 2018.

[Zontak *et al.*, 2013] Maria Zontak, Inbar Mosseri, and Michal Irani. Separating signal from noise using patch recurrence across scales. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1195–1202, 2013.