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/JBHI.2021.3092396, IEEE Journal of Biomedical and Health Informatics

The convolutional neural networks training with Channel-Selectivity for human activity recognition based on sensors

Wenbo Huang, Lei Zhang, Qi Teng, Chaoda Song and Jun He, *Member, IEEE*

***Abstract*—Recently, the state-of-the-art performance in various sensor based human activity recognition (HAR) tasks have been acquired by deep learning, which can extract automatically features from raw data. In order to obtain the best accuracy, many static layers have been always used to train deep neural networks, and their weight connectivity in network remains unchanged. Pursuing the best accuracy in mobile platforms with a very limited computational budget at millions of FLOPs is impractical. In this paper, we make use of shallow convolutional neural networks (CNNs) with channel-selectivity for the use of HAR. As we have known, it is for the first time to adopt channel-selectivity CNN for sensor based HAR tasks. We perform extensive experiments on 5 public benchmark HAR datasets consisting of UCI-HAR dataset, OPPORTUNITY dataset, UniMib-SHAR dataset, WISDM dataset, and PAMAP2 dataset. As a result, the channel-selectivity can achieve lower test errors than static layers. The existing performance of deep HAR can be further improved by the CNN with channel-selectivity without any extra cost.**

***Index Terms*—Sensor, convolutional neural networks, activity recognition, deep learning, channel-selectivity**

I. Introduction

With the rapid technical advance of smartphones and other wearable devices, a large variety of embedded inertial sensors like gyroscope and accelerometer enable researchers to collect human posture signal for activity monitoring. Wearable sensor based human activity recognition (HAR)[1][2] has turned into a new research hotspot with various real-world applications like health monitoring[3][4][5][6], sports tracking[7][8], smart homes[9], and game console designing[10] *etc*. Recently, deep learning[11] models with multiple layers have been built up to model high-level abstraction in sensor time series. Among different deep learning models especially attractive is convolutional neural network (CNN)[12] due to its specific architecture, which is able to automatically capture sensor patterns and their variations. It is well known that deep CNNs have achieved state-of-the-art performance across a variety of HAR tasks. However, compared with shallow networks, deep CNNs with many layers usually consume more computing resource which is impractical for wearable HAR scenario. There are continuous researches devoted to address sensor based HAR problems. In Computer Vision (CV) field, it has usually been assumed that that all convolutional layers are static and their weight connectivity within network remains unchanged during training stage. Several researchers have proposed to add more static layers in order to improve the performance of CNN. Typically, He *et al.* proposed a ResNet-101[13], which has 101 layers, thus increasing computational burden at millions of FLOPs. Pursuing the best accuracy in mobile platforms with a very limited computational budget is impractical. Various pruning or compressing techniques have been exploited to reduce memory or computational burden. For example, Liu *et al.* used network slimming[14] during training stage, which may undermine the generalization ability and classification accuracy of the model. Jeong *et al.* proposed selective allocation of channels[15] that dynamically increases the efficiency of CNN. To our knowledge, dynamic pruing techniques have seldom been exploited in HAR field. Therefore, designing a light-weight CNN for HAR that is able to achieve state-of-the-art performance is significant.

As we have known, when one recognizes an activity, only a few of channels are useful while other channels have little or even no contribution to recognition performance. In this paper, we take the inspiration from Jeong *et al.*[15] and use channel-selectivity to train CNNs for HAR applications. That is to say, during training process, each convolutional layer can pick out more important channels. As the training goes on, some of input channels for each convolution may contribute little or no to the output. Those channels waste many resources which should have been allocated in other channel training. The channel-selectivity has the ability to detect this kind of useless channels. The resources of these channels can be released to Top-K important channels. To sum up, the channel-selectivity is a dynamic pruning and re-wiring process which can improve the efficiency of CNN. In essence, the channel-selectivity simulates the way the human brain learns by hippocampus. During maintenance, new neurons are created and rewired via neuronal apoptosis or pruning[16].

The channel-selectivity consists of two parts. First of all, we use the Expected Channel Damage Matrix (ECDM)[15] to

The work was supported in part by the National Science Foundation of China under Grant 61203237 and the Industry-Academia Cooperation Innovation Fund Projection of Jiangsu Province under Grant BY2016001-02, and in part by the Natural Science Foundation of Jiangsu Province under grant BK20191371. *(Corresponding author: Lei Zhang.(e-mail: leizhang@njnu.edu.cn))*

Wenbo Huang, Lei Zhang, and Qi Teng are with School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing, 210023, China.

Chaoda Song is with School of Software, Zhengzhou University, Zhengzhou, 450002, China

Jun He is with the School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China.

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/JBHI.2021.3092396, IEEE Journal of Biomedical and Health Informatics

estimate the changes in output and judge which channel has little or no contribution and needs to be removed. During the training stage, the ECDM provides a safe yet effective way for removing or emphasizing channels. Besides, since using channel-selectivity could lead to unsatisfactory parameters recycling which may undermine the generalization, the spatial shifting[17][18][19] is imposed for more efficient parameter recycling. The results indicate that the spatial shifting is also enables to lead to larger kernel size for important channels.

We test this method on 5 datasets which can be publicly avaliable, including UCI-HAR dataset, OPPORTUNITY dataset, UniMib-SHAR dataset, WISDM dataset, and PAMAP2 dataset. By substituting traditional convolution for channel selective convolution, as well as adding spatial shifting, we can obtain the channel-selectivity CNN for HAR. The result shows the advantage of channel-selectivity in HAR. The performance improvement is analyzed in details, which enables us to draw a conclusion with regard to tpyical HAR applications.

The channel-selectivity brings a new and effective method via re-wiring idea. Compared with the existing methods, the channel-selectivity has several preponderances: (a) Extensive use: various kinds of deep HAR techniques can use it, (b) Compatibility: it can be seamlessly integrated into the existing deep HAR schemes that use wearable sensors, and (c) Equilibrium: it can achieve a trade-off between HAR accuracy improvement and computing budget easily. The structure of this paper is structured as follows. In Section II, we summarize the HAR's related works. The HAR with channel-selectivity is detailed in Section III. Section IV describes 5 HAR datasets used, experimental settings and main results. Ablation analysis and discussions are given in Section V, in which various training schemes and visualizing analysis across the channel importance are provided. In final Section VI, we conclude the paper.

II. Related works

In CV field, static layers have usually been used to train deep neural networks, which obviously lacks a plausible reason from biological viewpoint. Several researchers have used channel pruning technique to train deep models. Han *et al.*[20] have proposed channel pruning to avoid using too many useless parameters during training. He *et al.*[21] proposed a least absolute shrinkage and selection operator (LASSO) regression based channel selection in which least square reconstruction is used to accelerate very deep CNNs. Gao *et al.* proposed feature boosting and suppression (FBS)[22] to predictively amplify salient convolutional channels and skip unimportant ones at run-time. Ye *et al.*[23] proposed a channel pruning technique for accelerating the computations of deep CNNs. This technique focuses on direct simplification of the channel-to-channel computation graph of CNN. Although the channel-selectivity has many advantages such as accuracy improvement and biological plausibility, the channel-selective idea has rarly been exploited in the related HAR researches.

Recently, deep learning that is able to extract features automatically has attracted much attention in HAR community. For example, Jiang *et al.*[12] processed the raw sensor time series into 2-dimensional signal image. Then they use a 2D ConvNet with 2 layers to classify 2-dimensional images for inferring specific activity. Wang *et al.*[24] proposed a CNN that uses attention idea, in which weakly labeled sensor time series can be located and recognized. Teng *et al.*[25] imposed local loss into a 3-layer CNN, which is then used to classify raw sensor signals into various activities. Zeng *et al.*[26] converted each accelerometer's dimension into one channel that can be seen as a RGB image, in which convolutional layers and pooling layers can be in independent sensor dimension. However, sensor data unlike images, it not only has connection in spatial pixels but also has a time series relationship. Ordóñez *et al.*[27] firstly prosposed to combine long-term short memory (LSTM) and CNN, which can further improve accuracy of HAR via fusing multimodal sensors. Although static layers have been extensivly adopted in all above models, which could not obtain very good performance in recognition accuracy because of too many useless parameters caused by the channels with low contribution. Therefore, Zeng *et al.*[28] and Ma *et al.*[29] used attention in HAR to focus on the channels with more contribution for classification. For present, the static convolution has dominated the HAR research that uses deep learning. The channel-selevtivity mechanism has never been explored in related HAR literatures.

III. Model

Our research motivation is to use the channel-selectivity to upgrade standard convolutional layer in HAR applications. Selecting important channels with more contribution make more effective recycling of parameters feasible. In HAR scenario, what one has to be firstly dealed with is multi-channel sensor time series signals. Following the settings of the related literatures[15][17][18][19], there are two main operations that need to be imposed on these input channels:

1. Channel de-allocation(deallocation): Unnecessary channels are obstructed during training stage. Related parameters used in future computations are released.

2. Channel re-allocation(reallocation): Obstracted channels are replaced by Top-K important channels. The parameters of Top-K important channels are recycled and cover parameters of the obstructed channels.

In this way, these raw input may be converted to new channels with more important contributions after many repetitions. In other words, the important channels are retained. Fig.1 shows the structure of channel-selectivity in our model. The raw time series is collected from various sensors such as accelerometer, gyroscope and magnetometer. To maintain the continuity of time series signal, the data collected by these sensors is firstly segmented via using sliding window technique during the preprocessing stage. In previous method, the data windows are then fed into the standard CNN, which consists of N convolutional layers (Conv: convolutional layer,

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/JBHI.2021.3092396, IEEE Journal of Biomedical and Health Informatics

BN: Batch Normalization, ReLU: Rectified Linear Unit). When the channel-selectivity mechanism is used, the normal convolutional operation is replaced by the same number of channel-selective convolutional operation which consists of Dealloc: Channel Deallocation, Realloc: Channel Reallocation, and Spatial Shift: Spatial Shift operation. Here, Channel Deallocation and Channel Reallocation module is responbile for obstructing useless channels and then using other important channels to replace useless channels, while Spatial Shift can offset the loss of diversity caused by too many similar channels. Finally, the fully connected (FC) layer performs activity classification task. Fig.2 illustrates the two basic operations. The size of convolution kernels applied on sensor data along temporal dimension is $K\times1$, which is different from imagery data. The upper line represent I channels, each of which has different importantce. The more important channel has higher saturation. At the middle line, the unimportant channels are obstructed and the corresponding parameters are released during deallocation stage. At the bottom line, the Top-K important channels are copyed into the released areas. In order to find an effective and safe way to identify channels with low contribution to the output, we impose the Expected Channel Damage Matrix (ECDM)[15] (Section III-A). As the diversity of the convolutional layer should be maintained, we used spatial shifting[17][18][19] (Section III-B) to improve the generality ability of the model. In Section III-C, we provide more details of the training scheme using ECDM.

A. Expected Channel Damage Matrix

When one recognizes an activity, only a few of channels are useful while other channels have little even no contribution to recognition. Zeng *et al.*[28] selected useful channels of sensor data input by using the attention mechanism. In order to find useless layers, Zhang *et al.*[30] proposed channel shuffle within group convolutions. In the convolution Conv(X; W), X and W represent its input tensor and weight respectively, in which the input channel can be denoted by I and the output channel can be denoted by O. The height and width of the input are h and w. Due to various sensor modalities, the convolution kernel in CNN is only applied along the temporal dimension. Here the kernal size along temporal dimension is $K\times1$.

Expected channel damage matrix (ECDM)[15] can be used for gauging channels' expected functional difference. Setting W_i to 0 indicates the *i-th* channel is damaged or pruned, which needs to be blocked. That is to say, the ECDM measures the expected number of output's changes. For i=(1, 2, ... I), we compute the average of the expectation over temporal dimension to define $ECDM(X;W)$:

$$\begin{aligned} ECDM(X;W) &= Ex\left(X_i^{I\times h\times w};W_i^{I\times O\times K\times 1}\right)_i \\ &= \frac{1}{hw}\sum Ex\left[Conv(X;W)-Conv(X;W_i)\right]_{:,h,w} \end{aligned} \tag{1}$$

B. Selective Convolutional Layer

Recently, spatial shifting has an extraordinary performance in the model design of CNN. Jeon *et al.*[17], Dai *et al.*[18] and Wu *et al.*[19] used similar spatial shifting for visual recognition tasks. Actually, channel blocking and re-indexing are necessary in deallocation and reallocation[15] of channel-selectivity.

Here, g=(0, 1) represents gate variables. For i=(1, 2, ..., I), π=(1, 2, ..., I). The values of various π may be same. The input channel X is blocked if it has not important contribution. At this time g is equal to 0. One channel can be copied many times if this channel is important. However, simply copying an important channel may undermine classification accuracy and lose the diversity of the convolutional layer. Too many similar channels cause the corresponding weights to degenerate due to linear characteristics of the convolutional layer. In order to avoid the shortcoming, we use spatial shifting bias b in reallocation. The expression of *SelectChannel*[15] is the following equitions:

$$\begin{aligned} SelectChannel(X;g,\pi)_i &= g_i \times shift(X_{\pi_i},b_i) \\ b_i &= \left(b_i^h,b_i^w\right)\in R^2 \\ g_i &= \{0,1\} \\ \pi_i &= \{1,2,...,I\} \\ i &= \{1,...,I\} \end{aligned} \tag{2}$$

$Shift(X,b)$ represents the spatial shift operation of X, and we define $Shift(X,b)$ as:

$$\begin{aligned} shift(X,b)_x &= \sum_{n=1}^{H} X_n \times \max\left(0,1-|x-n+b^h|\right), \\ shift(X,b)_y &= \sum_{m=1}^{W} X_m \times \max\left(0,1-|y-m+b^w|\right). \end{aligned} \tag{3}$$

This trick can be used more efficiently with re-allocating parameters. During convolution process, it enhances these copied channels' diversity. Actually, Fig.3 is the sketch map of spatial shifting. This method is selectively able to expand the kernel size by recycling its parameters.

C. Training Scheme: Channel De/Re-allocation

In particular, deallocation and reallocation are designed to train the selective convolution with S = (W, g, π, b). Supposing deallocation, firstly giving a desired damaged level γ>0, we use a simple greedy algorithm to solve multi-dimensional knapsack problem (MKP)[15][31].

In comparison with other algorithm[32][33] that solve MKP, the greedy algorithm can yield a easier computation. Concerning the dimension of output, We normalize the ECDM, namely normalized-ECDM(nECDM)[15]:

$$nECDM(X;W)_{:,j} = \frac{|ECDM(X;W)_{:,j}|}{\sum_{i=1}^{I}|ECDM(X;W)_{:,j}|} \le \gamma \tag{4}$$

Assuming that j=(1, 2, ..., O), one can determine the channels that will be de-allocated according to the channel

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/JBHI.2021.3092396, IEEE Journal of Biomedical and Health Informatics

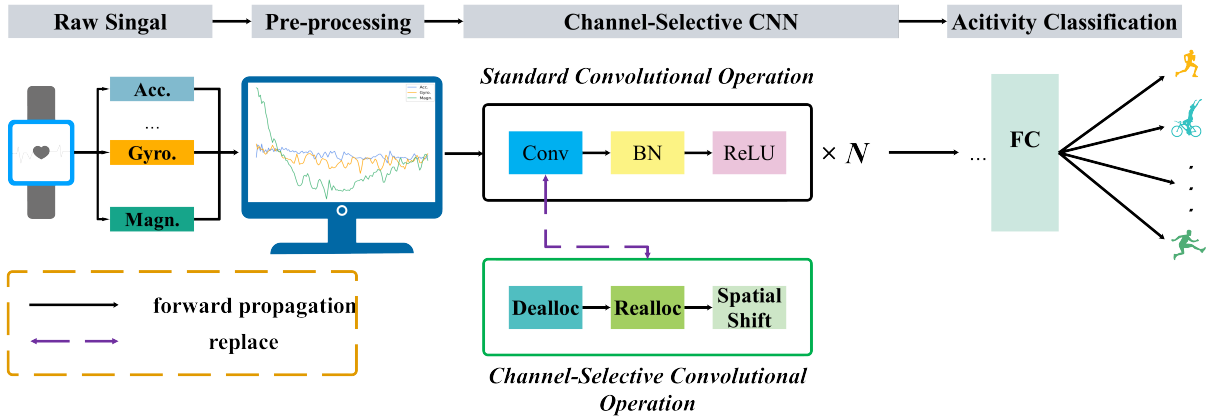


Fig. 1. Overview of the model for HAR with channel-selectivity. Acc., Gyro. and Magn. independently represent accelerometer, gyroscope and magnetometer. The data collected by these sensors will be sent to computer. On the computer screen, it is the plot of snesor data on time.

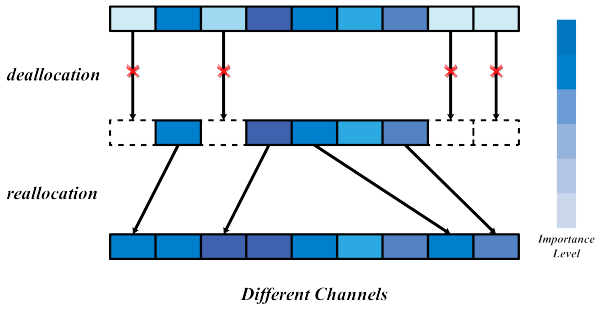


Fig. 2. Illustration of the producedures of channel deallocation and reallocation. The deep color of channels means the impotance is high.

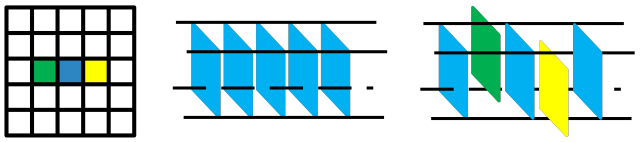


Fig. 3. spatial shifting has a effect of enlarging kernal.

of minimum $nECDM(X;W)$ iteratively on condition that the l^{∞}-norm of their sum of vector of $nECDM(X;W)$ is less than γ.

Supposing reallocation, we choose the Top-K most important channels according to $nECDM(X;W)$'s l^{∞}-norm. The channels that are currently de-allocated ($g = 0$) is occupied by the selected Top-K channels randomly. The corresponding parameters in W are set to zero when the *i-th* channel is de-allocated so that the operation does not damage the training. The maximum reallocation count $N_{\max}$ is also set to prevent too much reallocation for one feature.

Finally, the dynamic training scheme is built upon an existing Stochastic Gradient Descent (SGD)[34] method by simply calling deallocation or reallocation on demand. On the other hand, when we train W via SGD, deallocation and reallocation will update the other parameters of S: deallocation is used to update g, while reallocation is for the renewals of b, W, g and π.

In Section IV, our results reveal that the channel-selectivity can further improve classification performance. We also do a series of ablation studies in Section V to compare different training schemes and provide visualizing analysis across the channel importance level using wearable sensors.

IV. Experiments and Results

On various HAR datasets (UCI-HAR dataset, UniMib-SHAR dataset, OPPORTUNITY dataset, PAMAP2 dataset and WISDM dataset), we conduct our experiments. We use two baselines *i.e.* CNN and ResNet. The CNN has six convolutional layers. A fully connected layer is also used. The CNN and ResNet are used as baselines to evaluate the performance improvement caused by the channel-selectivity. After each activation function, BN is imposed. In the case of OPPORTUNITY and PAMAP2, we use selective convolution layer at the beginning since their input channels have more than one channel. For different datasets, we set different hyperparameters(γ, K and $N_{\max}$) to train the model.

For activity recognition tasks that use deep learning, data segmentation is an important stage[25][35]. Segmenting data stream is necessary for preprocessing. During data segmentation, a mainstream approach is sliding window technique. Recognition system's practical demands actually determine window size. We can find a specific window to improve the quality of one individual activity's recognition. In a general way, large windows are suitful for recognizing complex activities while small window size is good at recognizing faster activity. Banos *et al.*[45] recently investigated the effect of window size when classifying various activities. However, the best window size for deep learning is still unknown. Thus we apply the same window size of previous successful cases[25][35]. We choose SGD optimization method as the optimizer to train our models. According to different datasets, we set the initial learning rate. The experiments are implemented in Pytorch [46] deep learning framework. All experiments are conducted on a machine[25][35] with a 11GB NVIDIA 2080ti GPU, Intel i7 6850k CPU and 64 GB memory.

In Table.I, we summarize various attributes of 5 datasets used. At the same time, the setting of sliding windows in the

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/JBHI.2021.3092396, IEEE Journal of Biomedical and Health Informatics

TABLE I
SIMPLE DESCRIPTION OF DATASETS

Attribute \ Dataset	UCI-HAR	OPPORTUNITY	UniMib-SHAR	WISDM	PAMAP2
Sampling Rates	50Hz	30Hz	50Hz	20Hz	100Hz
Number of Categories	6	17	17	6	12
Proportion of Training Data	70%	70%	70%	70%	80%
Proportion of Testing Data	30%	30%	30%	30%	20%
Sliding Window Size	128	64	151	200	512
Sliding Window Step	64	32	76	20	256
Overlap Rates	50%	50%	50%	10%	50%

TABLE II
SIMPLE DESCRIPTION OF THE BASELINE CNN AND RESNET.

Simple Description \ Dataset	UCI-HAR	OPPORTUNITY	UniMib-SHAR	WISDM	PAMAP2
Layer1	C(64)	C(64)	C(64)	C(128)	C(128)
Layer2	C(64)	C(64)	C(64)	C(128)	C(128)
Layer3	C(128)	C(256)	C(256)	C(256)	C(256)
Layer4	C(128)	C(256)	C(256)	C(256)	C(256)
Layer5	C(256)	C(384)	C(384)	C(384)	C(384)
Layer6	C(256)	C(384)	C(384)	C(384)	C(384)
FC	✓	✓	✓	✓	✓
Softmax	✓	✓	✓	✓	✓
Training time(epoch)	200	200	200	200	200
Batch size	64	1024	64	64	128
Learning rate	0.001	0.001	0.001	0.001	0.001

TABLE III
ACCURACY(%) PERFORMANCE OF MODELS ON VARIOUS DATASETS

Model + Method \ Dataset	UCI-HAR	OPPORTUNITY	UniMib-SHAR	WISDM	PAMAP2
Baseline	96.12&0.33M	77.41&1.37M	75.97&1.55M	97.01&1.50M	90.57&0.86M
Baseline + SelectConv	**96.77**&0.33M	**79.67**&1.37M	**77.26**&1.55M	**97.44**&1.50M	**91.46**&0.86M
ResNet	96.33&0.84M	79.09&4.27M	76.93&3.62M	98.22&3.72M	91.53&3.79M
ResNet + SelectConv	**97.28**&0.84M	**82.36**&4.27M	**78.25**&3.62M	**98.52**&3.72M	**94.33**&3.79M
Other Researchers' Results	**96.98**[25]	81.00[25]	**78.07**[25]	97.50[25]	93.03[25]
	96.90[35]	76.83[26]	74.46[35]	96.90[35]	93.50[35]
	95.75[36]	**82.30**[37]	74.66[38]	93.32[39]	**93.70**[40]
	95.18[12]	75.74[27]	77.27[41]	97.51[42]	86.00[43]
	96.37[39]	74.50[40]	-	**98.20**[44]	89.96[28]

stage of pre-processing is also introduced.

The simple description of the baseline CNN and ResNet on different datasets can be seen in Table.II. A convolutional layer has L_s feature maps can be represented by C(L_s). We set the learning rate to decay exponentially. The ResNet has the same input and output with the baseline CNN.

1) UCI-HAR dataset[47]: The dataset was collected by 30 volunteers. Their ages range from 19 to 48 years. Each person wore a Sumsung Galaxy S2 smartphone on his waist and executed 6 different activities.

On the basis of the baseline CNN and ResNet, we add the channel-selectivity into each convolution layer. The test error curves are shown in Fig.4. We set the γ to 0.0005. At the first layer of CNN, since the input feature map has only one channel, we do not use the channel-selectivity. At the second layer, the K is set to 8 and the $N_{\max}$ is set to 16. At the third and fourth layer, the K is set to 16 and the $N_{\max}$ is set to 32. At the fifth and sixth layer, the K is set to 32 and the $N_{\max}$ is set to 64. It can be seen in Table.III that our accuracy is 97.28%. It obtains 2.1% and 1.53% performance gain over Ronao *et al.* (95.75%)[36] and Jiang *et al.* (95.18%)[12] using CNN. It

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/JBHI.2021.3092396, IEEE Journal of Biomedical and Health Informatics

also yields 0.38% improvement over Tang *et al.* (96.90%)[35]. When compared with Ignatov *et al.*'s results (96.37%)[39], our method achieves 0.91% accuracy improvement.

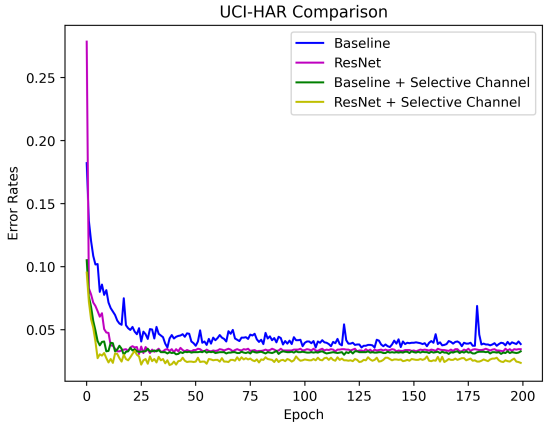


Fig. 4. UCI-HAR dataset's test error on different models.

2) OPPORTUNITY dataset[48]: The OPPORTUNITY dataset has been collected in a sensor-rich environment which consists of 15 wired and wireless network sensor systems. In this environment, the volunteers with on-body sensors performed their daily behavior in morning in order to collect 17 kinds of activities.

On the basis of baseline CNN and ResNet, we add the channel-selectivity into each convolution layer. The error rates are shown in Fig.5. We set the γ to 0.0005. The K and the $N_{\max}$ are both set to 16 at the first layer and second layer. Then, the K is set to 32 and the $N_{\max}$ is set to 32 at the third layer and fourth layer. Finally, at the fifth layer and sixth layer, the K is set to 64 and the $N_{\max}$ is also set to 64. The result can be seen in Table.III, in which the channel-selectivity can improve both ResNet and CNN. According to Table.III, our method (82.36%) outperforms Zhang *et al.*'s result (82.30%)[37] by 0.06% while using far less parameters. Our result that uses channel-selectivity in ResNet surpasses Zeng *et al.* (76.83%)[26], Hammerla *et al.* (74.50%)[40], Ordóñez *et al.* (75.74%)[27] and Teng *et al.*'s[25] result (81.00%).

3) UniMib-SHAR dataset[49]: The whole dataset was gathered by smartphone based on Android operating system. It can be used for detecting and recognizing fall activities. The dataset has 11,771 human activity samples. The 30 subjects within 18 and 60 years old joined in this collection process.

We add the channel selective submodule to CNN and ResNet. The γ is adjusted to 0.0005. Since the input feature map has only one channel, the channel-selectivity is useless. Thus we use the normal convolution at the first layer. The K is 8 and $N_{\max}$ is 16 at second layer. At the next two layers, the K is adjusted to 16 while the $N_{\max}$ is adjusted to 32. Finally, the K is set to 32 and the $N_{\max}$ is adjusted to 64 at the final two layers. The error rates is shown in Fig.6 and the experiment result is shown in Table.III. As we have known, Yang et al's result[41] was 77.27% via dynamic fusion method. Our result is 78.25%, which obtains 3.59% and 3.79% accuracy gain over Li *et al.* (74.66%)[38] and Tang *et al.* (74.46%)[35].

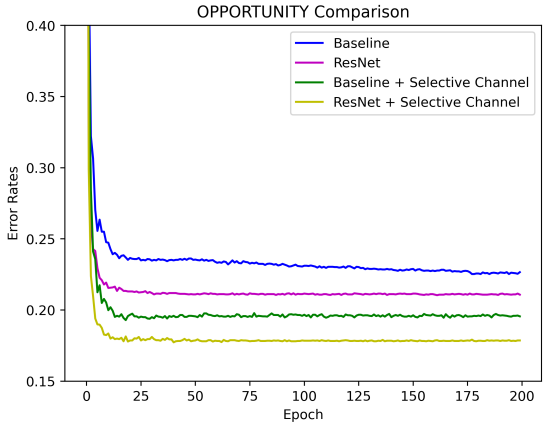


Fig. 5. OPPORTUNITY dataset's test error on different models.

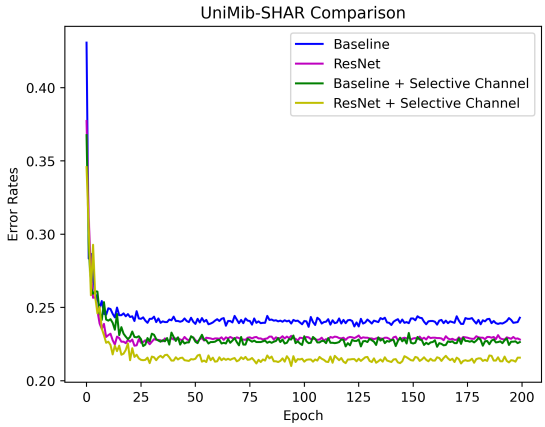


Fig. 6. UniMib-SHAR dataset's test error on different models.

4) WISDM dataset[44]: The WISDM group collected the WISDM dataset. They used smartphones (Android operating system) with accelerometer sensors to sample sensor time series. Each person attached smartphone in front leg pockets and performed 6 kinds of activities under supervision.

The channel-selectivity is used to replace the normal convolution layer. The test error curves are shown in Fig.7. We do not use the channel selective convolution layer at the first since the input feature map has only one channel. The γ is adjusted to 0.0001. At the second layer, the K is set to 16 and the $N_{\max}$ is set to 32. At the third layer and fourth layer, the K is set to 16 and the $N_{\max}$ is set to 32. The K is set to 32 and the $N_{\max}$ is set to 64 at the fifth layer and sixth layer. Table.III shows the experimental results. As we have known, the most high performance reported in the past for this dataset is 98.23% (Alsheikh *et al.*[42]), which used greedy layer-wise training based on deep belief networks. Our method (98.52%) is still able to produce an accuracy improvement on WISDM dataset, beating Ravi et al's result (98.2%) by 0.32% while consuming

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/JBHI.2021.3092396, IEEE Journal of Biomedical and Health Informatics

fewer parameters. Our result with channel-selectivity surpasses these results including Ignatov *et al.* (93.32%)[39], Teng *et al.* (96.90%)[25] and Tang *et al.* (97.50%)[35].

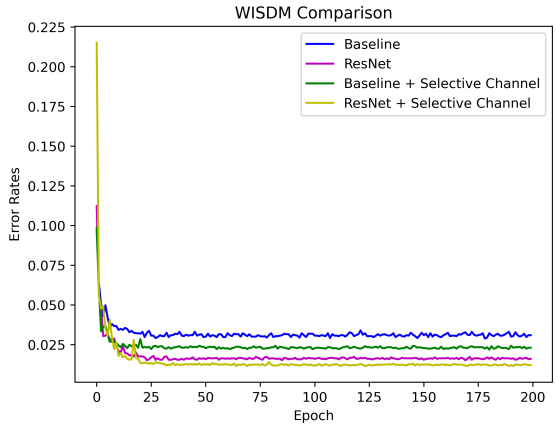


Fig. 7. WISDM dataset's test error on different models.

5) PAMAP2 dataset[50]: The PAMAP2 dataset can be used for monitoring physical activities. It was performed by 9 subjects, who wore three inertial measurement units (IMUs). Subjects' heart rate was also monitored by heart rate monitors. Researchers can use this dataset for strength estimation and activity recognition. Various tasks such as classification, feature extraction, data processing and segmentation can also be developed and evaluated on this dataset.

We replace the normal convolution layer by the channel selective convolution layer. In Fig.8, we show the curves of test error. The γ is set to 0.0005 during the whole training. At the first two layers, we adjust the K and $N_{\max}$ to 16 and 32 respectively. Then, at the subsequent two layers, the K and $N_{\max}$ are equal to 16 and 32 respectively. The final two layers have the same K while the $N_{\max}$ is adjusted to 48. The result can be seen in Table.III. Our result trained with the channel-selectivity methods is 94.33% , which achieves 0.63% improvement over the best result published by Hammerla *et al.* (93.70%)[40]. Our result also surpasses these results including Khan *et al.* (86.00%)[43], Tang *et al.* (93.50%)[35] and Zeng *et al.* (89.96%)[28].

We also compare the channel-selectivity CNN with our previous local loss method and its corresponding Lego variant. The resuts are shown in Table.III. It shows that channel-selevtivity is able to outperforms or match our previous techniques[25][35]. To be specific, the channel-selectivity method outperform our local-loss method 0.38%, 1.36%, 0.18%, 1.02%, 1.3% respectively on UCI-HAR dataset, OPPORTUNITY dataset, WISDM dataset, UniMib-SHAR dataset and PAMAP2 dataset.

V. Analysis and Discussions

In this section, we use the baseline CNN with 6 layers. Each of them uses the channel selectivity mechanism. We try to

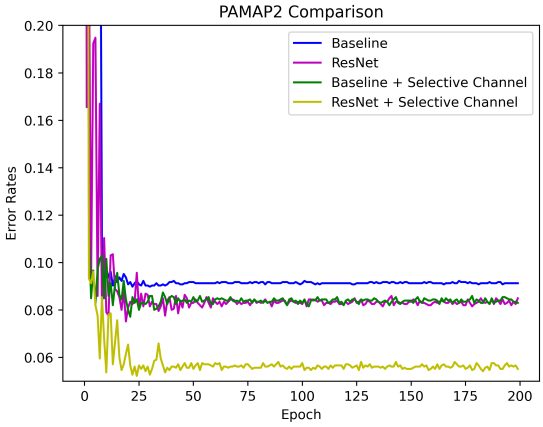


Fig. 8. PAMAP2 dataset's test error on different models.

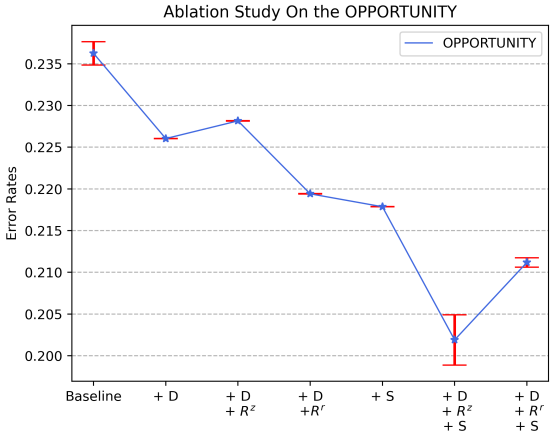


Fig. 9. OPPORTUNITY dataset's error rates on five CNN models with different re/de-allocation scheme.

find which part of the channel-selectivity works. We compare five different channel re/de-allocation schemes:

The Zero channel-selectivity(+D+R^z+S): Setting the corresponding convolution weights to 0 when one channel is re-allocated. Spatial shifting is used too.

The Random channel-selectivity(+D+R^r+S): Setting the corresponding convolution weights randomly when one channel is re-allocated. Spatial shifting is used at the same time.

Zero re-initialization(+D+R^z): Setting the corresponding convolution weights to 0 when one channel is re-allocated. Spatial shifting is not used here.

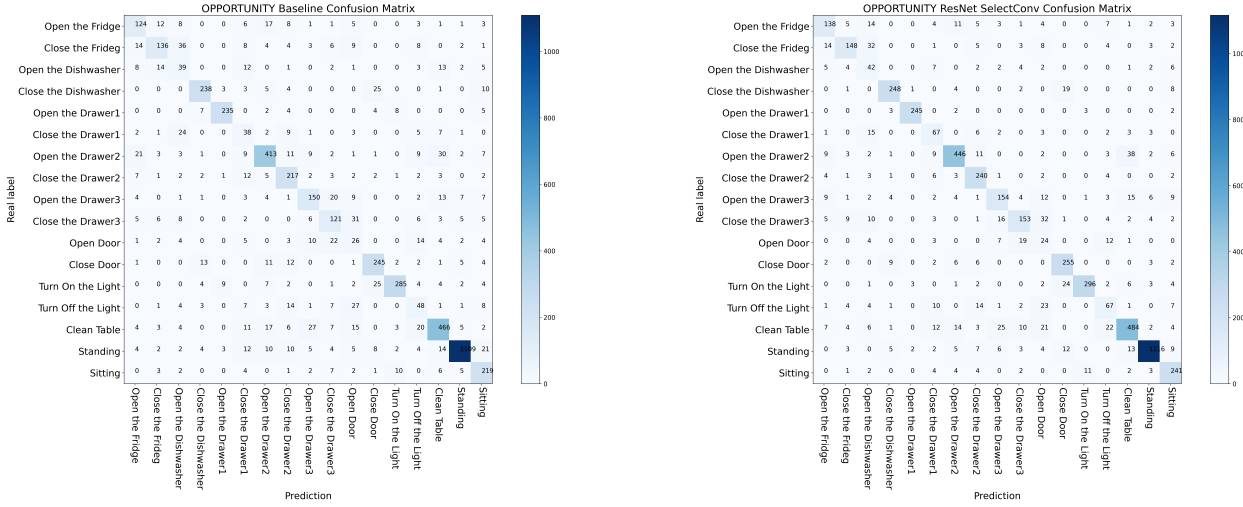
Random re-initialization(+D+R^r): Setting the corresponding convolution weights randomly when one channel is re-allocated. Spatial shifting is not used.

De-allocation only(+D): Only deallocation is used.

Shift only(+S)[17][18][19]: All channels use spatial shift without re/deallocation.

We calculate the mean value and variance of five different CNN models on OPPORTUNITY dataset with different re/de-allocation scheme. In Fig.9, it can be seen that +D+R^z+S have the best performance among all re/de-allocation schemes. The results indicate that using de/re-allocation alone is not

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/JBHI.2021.3092396, IEEE Journal of Biomedical and Health Informatics



(a) Baseline

(b) ResNet + Channel-Selectivity

Fig. 10. OPPORTUNITY dataset's Confusion Matrix on different methods

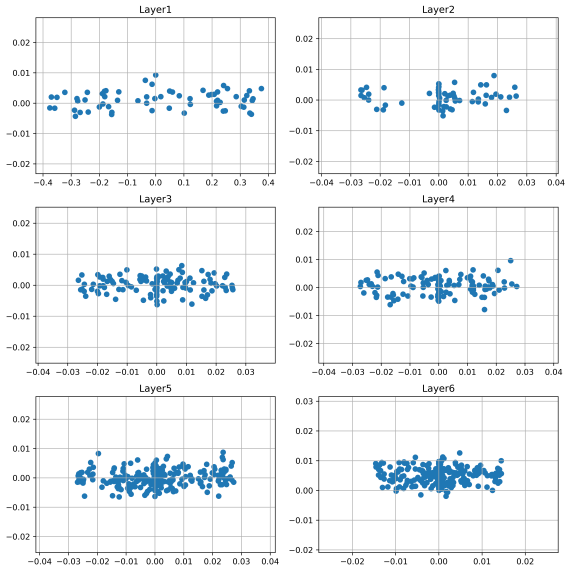
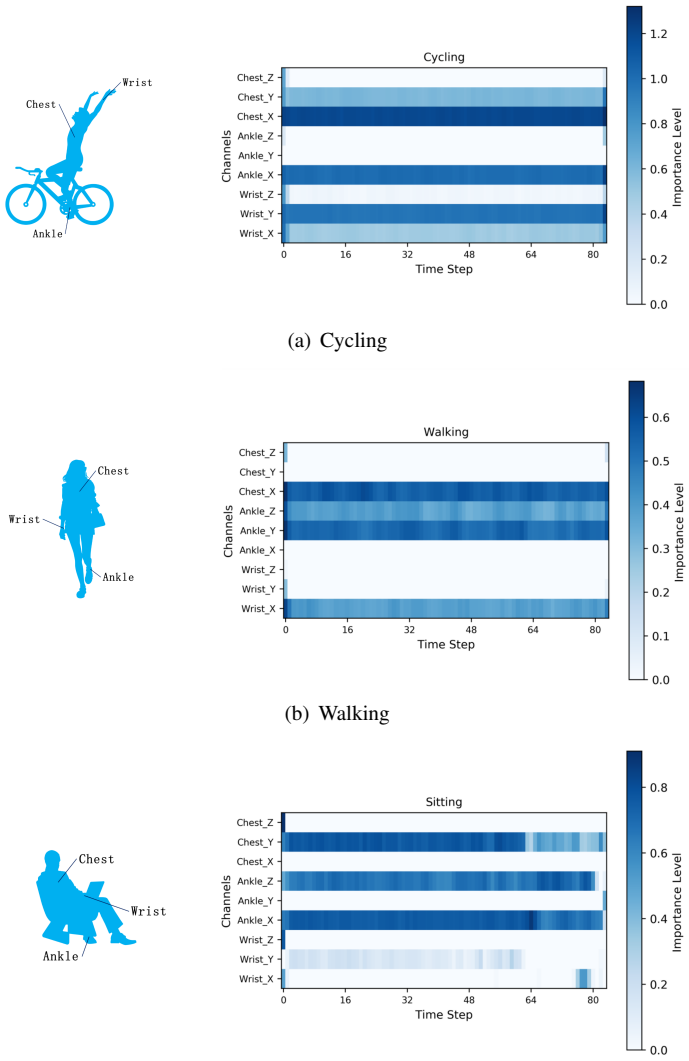


Fig. 11. Spatial shift biases tested on OPPORTUNITY dataset

enough.

In Fig.10, we compute confusion matrices on OPPORTUNITY dataset. When the channel-selectivity is integrated into ResNet, which yields much lower test errors than that of the baseline CNN.

The spatial shift technique has been explained in Section III-B, which is able to enlarge kernel size and effectively preserve the generality ability of model. Fig.11 shows how the spatial shift works within the channel-selectivity CNN on OPPORTUNITY dataset. What we can see is that the biases caused by spatial shift have an obvious tendency to align at the earlier layers. At the final layers, the spatial shift tend to enlarge the amount of new information that comes from previous channels. On the other hand, the biases at the final layers demonstrate less tendency to align, but a larger diversity on the value, which suggests larger kernel sizes.



(a) Cycling

(b) Walking

(c) Sitting

Fig. 12. Channel significance's visualization

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/JBHI.2021.3092396, IEEE Journal of Biomedical and Health Informatics

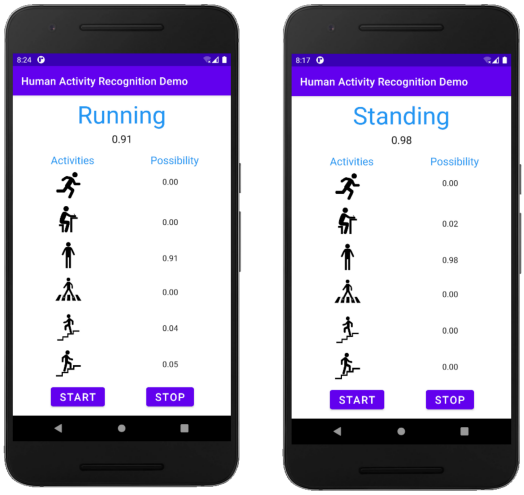


Fig. 13. Demo Application on mobile phone (Google Nexus 6)

TABLE IV
ACTUAL INFERENCE TIME ON SMARTPHONE

Method	Actual Inference Time (ms)
Baseline CNN	193(±20)
Channel-Selectivity	195(±25)

By visually analyzing the influence of different sensors placed on various human body's parts, we further evaluate the effect of channel-selectivity mechanism in multi-modal HAR scenario. We conduct experiments on a multimodal sensor dataset, *i.e.*, PAMAP2. The channel-selectivity consciously is able to automatically learn different sensor's precedence rather than equally handle all sensor modalities. Therefore, the channel-selectivity yields excellent results in deep feature fusion and multimodal HAR tasks. The results of Fig.12 are plausible and can be easily understood according to people's daily intuition.

Finally, in order to demonstrate the superiority of our method, we run the channel-selectivity model on real mobile devices. Following the guidence of the APP[51] which has released source code, we obtain the actual inference time. In Fig.13, we provide several screenshots of the practical test in APP. On WISDM datast, we train the channel-selectivity and normal CNN respectively. We aim to construct an application which can be run on Android operation system. The trained models are transferred into .pb files. The APP is operated on a Google Nexus 6 phone which is equipped with an Android 11.0 operation system. Table.IV shows the comparasion of inference time. In the test, there is nearly no any difference between channel-selectivity and normal CNN according to inference speed.

VI. CONCLUSION

In this paper, the channel-selectivity idea is for the first time adopted in HAR scenario. The CNN can be trained dynamically with the channel-selectivity operation combined by de/re-allocation and spatial shift. On 5 public HAR datasets including UCI-HAR dataset, OPPORTUNITY dataset, UniMib-SHAR dataset, WISDM dataset, and PAMAP2 dataset, we perform extensive experiments. We construct baseline CNN and ResNet for each dataset. The result is comparable with the previous results reported in relevant literatures. As mentioned above, the channel-selectivity can lead to a dynamic and efficient training. Since more important or decisive channels are selected, the channel-selectivity can obtain lower test errors in HAR. Experimental results show that the channel-selectivity can consistently improve test errors of baselines. There is not a significant increase in terms of computation overhead.

For standard convolution, many sensor channels almost have no contribute to output, while these channels still occupy too many resources. On the other hand, some inputs channels have more important or decisive influence to output. Visualization of channel importance of various activities in PAMAP2 dataset is shown in Fig.12. Compared to standard convolution, the channel-selectivity module can find these channels of low contribution via ECDM. Then the resources of these channels with low contribution are reallocated for the Top-K most important or decisive channels. Under this circumstance, the important channels can be duplicated multiple times. After the allocation, the input channels will all have contribution to the output. However, due to the linearity of convolutional layer, naively copying one channel in the allocation process does not provide any benefit. It is well known that the corresponding model will degenerate if two input channels are completely identical. Therefore, to avoid this shortcoming, the spatial shifting technique is applied to enhance the diversity on these copied channels, which can improve the generality ability for the channel-selectivity CNN. Five different channel re/de-allocation schemes with/without spatial shifting is compared in Fig.9 and Fig.11. To sum up, without any extra cost such as number of paramerters, our experimental results indicates that the channel-selectivity method can lead to a significant performance improvement on various wearable HAR tasks (Table.III). De/re-allocation can reuse memory. In HAR tasks based on wearable sensors, the reuse of memory is very useful. The channel-selectivity has a plausible explanation in biology. Its compelling performance on the benchmark HAR datasets is also verified by our results. We believe that channel-selectivity will open a new direction of training CNNs for HAR.

REFERENCES

[1] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[2] M. Zhang and A. A. Sawchuk, "Human daily activity recognition with sparse representation using wearable sensors," *IEEE journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 553–560, 2013.

[3] Y.-J. Hong, I.-J. Kim, S. C. Ahn, and H.-G. Kim, "Mobile health monitoring system based on activity recognition using accelerometer," *Simulation Modelling Practice and Theory*, vol. 18, no. 4, pp. 446–455, 2010.

[4] N. Alshurafa, W. Xu, J. J. Liu, M.-C. Huang, B. Mortazavi, C. K. Roberts, and M. Sarrafzadeh, "Designing a robust activity recognition framework for health and exergaming using wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1636–1646, 2013.

[5] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Oehler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 140–148, 2014.

[6] H. Monkaresi, R. A. Calvo, and H. Yan, "A machine learning approach to improve contactless heart rate monitoring using a webcam," *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1153–1160, 2013.

[7] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *computers*, vol. 2, no. 2, pp. 88–131, 2013.

[8] 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, pp. 1–33, 2014.

[9] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[10] P. Rashidi and D. J. Cook, "Keeping the resident in the loop: Adapting the smart home to the user." *IEEE Trans. Systems, Man, and Cybernetics, Part A*, vol. 39, no. 5, pp. 949–959, 2009.

[11] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[12] 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*, 2015, pp. 1307–1310.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2736–2744.

[15] J. Jeong and J. Shin, "Training cnns with selective allocation of channels," *arXiv preprint arXiv:1905.04509*, 2019.

[16] A. Sahay, K. N. Scobie, A. S. Hill, C. M. O'Carroll, M. A. Kheirbek, N. S. Burghardt, A. A. Fenton, A. Dranovsky, and R. Hen, "Increasing adult hippocampal neurogenesis is sufficient to improve pattern separation," *Nature*, vol. 472, no. 7344, pp. 466–470, 2011.

[17] Y. Jeon and J. Kim, "Active convolution: Learning the shape of convolution for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4201–4209.

[18] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[19] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer, "Shift: A zero flop, zero parameter alternative to spatial convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9127–9135.

[20] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[21] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1389–1397.

[22] X. Gao, Y. Zhao, Ł. Dudziak, R. Mullins, and C.-z. Xu, "Dynamic channel pruning: Feature boosting and suppression," *arXiv preprint arXiv:1810.05331*, 2018.

[23] J. Ye, X. Lu, Z. Lin, and J. Z. Wang, "Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers," *arXiv preprint arXiv:1802.00124*, 2018.

[24] K. Wang, J. He, and L. Zhang, "Attention-based convolutional neural network for weakly labeled human activities recognition with wearable sensors," *IEEE Sensors Journal*, vol. 19, no. 17, pp. 7598–7604, 2019.

[25] Q. Teng, K. Wang, L. Zhang, and J. He, "The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition," *IEEE Sensors Journal*, vol. 20, no. 13, pp. 7265–7274, 2020.

[26] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *6th International Conference on Mobile Computing, Applications and Services*. IEEE, 2014, pp. 197–205.

[27] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[28] M. Zeng, H. Gao, T. Yu, O. J. Mengshoel, H. Langseth, I. Lane, and X. Liu, "Understanding and improving recurrent networks for human activity recognition by continuous attention," in *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, 2018, pp. 56–63.

[29] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, "Attnsense: Multi-level attention mechanism for multimodal human activity recognition." in *IJCAI*, 2019, pp. 3109–3115.

[30] 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.

[31] H. Kellerer, U. Pferschy, and D. Pisinger, "Multidimensional knapsack problems," in *Knapsack problems*. Springer, 2004, pp. 235–283.

[32] M. Vasquez and Y. Vimont, "Improved results on the 0–1 multidimensional knapsack problem," *European Journal of Operational Research*, vol. 165, no. 1, pp. 70–81, 2005.

[33] G. R. Raidl and J. Gottlieb, "Empirical analysis of locality, heritability and heuristic bias in evolutionary algorithms: A case study for the multidimensional knapsack problem," *Evolutionary computation*, vol. 13, no. 4, pp. 441–475, 2005.

[34] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[35] Y. Tang, Q. Teng, L. Zhang, F. Min, and J. He, "Layer-wise training convolutional neural networks with smaller filters for human activity recognition using wearable sensors," *IEEE Sensors Journal*, 2020.

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

[37] L. Zhang, X. Wu, and D. Luo, "Recognizing human activities from raw accelerometer data using deep neural networks," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 865–870.

[38] F. Li, K. Shirahama, M. A. Nisar, L. Köping, and M. Grzegorzek, "Comparison of feature learning methods for human activity recognition using wearable sensors," *Sensors*, vol. 18, no. 2, p. 679, 2018.

[39] A. Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Applied Soft Computing*, vol. 62, pp. 915–922, 2018.

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

[41] Z. Yang, O. I. Raymond, C. Zhang, Y. Wan, and J. Long, "Dfternet: Towards 2-bit dynamic fusion networks for accurate human activity recognition," *IEEE Access*, vol. 6, pp. 56 750–56 764, 2018.

[42] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, "Deep activity recognition models with triaxial accelerometers," in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[43] A. Khan, N. Hammerla, S. Mellor, and T. Plötz, "Optimising sampling rates for accelerometer-based human activity recognition," *Pattern Recognition Letters*, vol. 73, pp. 33–40, 2016.

[44] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, "Deep learning for human activity recognition: A resource efficient implementation on low-power devices," in *2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN)*. IEEE, 2016, pp. 71–76.

[45] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474–6499, 2014.

[46] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

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

[48] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha *et al.*, "Collecting complex activity datasets in highly rich networked sensor environments," in *2010 Seventh international conference on networked sensing systems (INSS)*. IEEE, 2010, pp. 233–240.

[49] D. Micucci, M. Mobilio, and P. Napoletano, "Unimib shar: A dataset for human activity recognition using acceleration data from smartphones," *Applied Sciences*, vol. 7, no. 10, p. 1101, 2017.

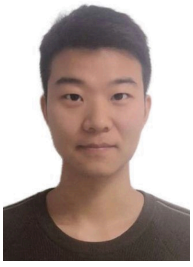
[50] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *2012 16th International Symposium on Wearable Computers*. IEEE, 2012, pp. 108–109.

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/JBHI.2021.3092396, IEEE Journal of Biomedical and Health Informatics

[51] D. Singh, E. Merdivan, I. Psychoula, J. Kropf, S. Hanke, M. Geist, and A. Holzinger, "Human activity recognition using recurrent neural networks," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2017, pp. 267–274.

Wenbo Huang received the B.S. degree from Nanjing Tech University, Nanjing, China, in 2019. He is currently pursuing the M.S. degree with Nanjing Normal University. His research interests include activity recognition, computer vision, and machine learning.

Lei Zhang received the B.Sc. degree in computer science from Zhengzhou Universitiy, China, and the M.S. degree in pattern recognition and intelligent system from Chinese Academy of Sciences, China, received the Ph.D. degree from Southeast University,China, in 2011. He was a Research Fellow with IPAM, UCLA, in 2008. He is currently an Associate Professor with the School of Electrical and Automation Engineering, Nanjing Normal University. His research interests include machine learning, human activity recognition and computer vision.



Qi Teng received the B.S. degree from Henan University of Engineering, Zhengzhou, China, in 2018. He is currently pursuing the M.S. degree with Nanjing Normal University. His research interests include activity recognition, computer vision, and machine learning.

Chaoda Song is currently pursuing the B.S degree with Zhengzhou University. His research interests include activity recognition, computer vision, and machine learning.

Jun He received the Ph.D. degree from Southeast University, Nanjing, China, in 2009. He was a Research Fellow with IPAM, UCLA, in 2008. From 2010 to 2011, he was a Post-Doctoral Research Associate with the Chinese University of Hong Kong. He is currently an Associate Professor with the School of Electronic and Information Engineering, Nanjing University of Information Science and Technology. His main research is in the areas of machine learning, computer vision, and optimization methods. In particular, he is interested in the applications of weakly supervised learning via reinforcement learning methods.