Visualizing Inertial Data For Wearable Sensor Based Daily Life Activity Recognition Using Convolutional Neural Network*

Thien Huynh-The[1], Cam-Hao Hua[2] and Dong-Seong Kim[1]

***Abstract*— Nowadays human activity recognition (HAR) plays an crucial role in the healthcare and wellness domains, for example, HAR contributes to context-aware systems like elder home assistance and care as a core technology. Despite promising performance in terms of recognition accuracy achieved by the advancement of machine learning for classification tasks, most of the existing HAR approaches, which adopt low-level handcrafted features, cannot completely deal with practical activities. Therefore, in this paper, we present an efficient wearable sensor based activity recognition method that allows encoding inertial data into color image data for learning highly discriminative features by convolutional neural networks (CNNs). The proposed data encoding technique converts tri-axial samples to color pixels and then arranges them for image-formed representation. Our method reaches the recognition accuracy of over 95% on two challenging activities datasets and further outperforms other deep learning-based HAR approaches.**

I. INTRODUCTION

In the last decade, human activity recognition (HAR) has attracted considerable attention from the healthcare society due to playing an important role of activity monitoring in several health and wellness applications [1], for example, heart failure detection and cardiovascular disease recognition. The knowledge of daily physical activity is useful for blood pressure control and so far helps preventing dangerous heart diseases like myocardial infarction. Fundamentally, activity recognition is categorized into two groups: visual-based and wearable sensor-based. Despite of rich information, visual-based HAR has such essential limitations as illumination change, occlusion, and personal privacy issue [2]. Thanks to the ability of continuous monitoring physical activities without the influence of surrounding objects if compared with visual-based HAR, wearable sensor-based approaches have been widely adopted in the health-wellness analysis and diagnosis. A large number of wearable sensor-based HAR works have been conducted with the remarkable improvement of performance by the advancements of sensor technology and machine learning technique. However, the most of them, based on low-level handcrafted features and classical classification methods, are usually fragile when dealing with complex activities in the practical environment [3].

Several conventional HAR methods have exploited descriptive statistic features (e.g., mean, median, variance, standard deviation) [4] or transformed-domain features (e.g., discrete cosine transform, fast Fourier transform, and Wavelet transform) [5] coupled with traditional classification techniques (e.g., Bayesian decision making, rule-based algorithm, least-squares method, k-nearest neighbor, support vector machine, and artificial neural networks) [6]. To improve the recognition accuracy of complex activities, a hierarchical classification architecture was proposed for well working with multi-sensors [8], [9], instead of a single classifier for each sensory channel (e.g., accelerometer and gyroscope). Fusion techniques [10], including data-level and decision-level, were further studied for handling the activity recognition from multiple sensory sources. Recently, recurrent neural networks (RNNs) [11]–[13] and long short-term memory (LSTM) network (a.k.a., an advancement of RNNs) [14]–[16] have been considered for automatically learning discriminative features from raw sensory data. Feature extraction and selection processes are unnecessary (i.e., they are performed inside a deep network during the training stage) to reduce the complexity of a general recognition workflow. Additionally, RNNs and LSTM network models are extraordinarily competent in capturing the temporal correlations of sequential data. Lately, convolutional neural networks (CNNs), widely used for classification task in image processing and computer vision, are developed for recognizing daily activities using inertial sensory data [17], [18]. Compared with RNN-based activity recognition methods which directly process the input of raw sequential data, CNN-based approaches require to represent wearable sensory data by image in advance. For example, Zeng et al. [17] simply plotted three signal images corresponding to the accelerometer data of three axes in the time domain before feeding to a deep CNN for learning recognition model, however, this inertial-data-to-image representation scheme is ineffective due to losing detail information and wasting computational resource for blank area. Although conducting significant improvement of accuracy in comparison with classical HAR approaches, current deep learning-based methods cannot thoroughly deal with the challenges of multiple sensors-based activity modeling and high-level feature learning.

In this paper, we propose an efficient wearable sensor-based activity recognition approach by exploiting a novel data encoding mechanism coupled with a pre-trained CNNs model. In order to fit with the input of CNNs models (i.e., image or video) for classification tasks, the inertial sensory data from accelerometer and gyroscope are encoded to image-formed data, wherein each discrete-time sequential sample, presented by a set of three values corresponding

[1]ICT Convergence Research Center, Kumoh National Institute of Technology, Gumi, South Korea thienht@kumoh.ac.kr, dskim@kumoh.ac.kr

[2]Computer Science and Engineering, Kyung Hee University, Yongin, South Korea hao.hua@uclab.ac.kr

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

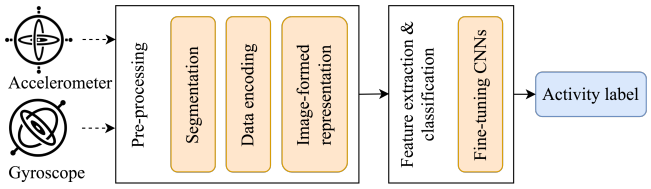
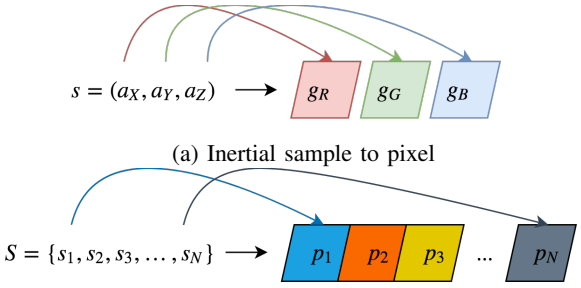


Fig. 1: The overall workflow of wearable sensor-based activity recognition with data encoding and CNNs fine-tuning.



(a) Inertial sample to pixel

(b) Inertial sequence to pixel array

Fig. 2: The illustration of encoding inertial data to image-formed data.

to three axes, is transformed to a color pixel. Accordingly, a segmented activity sequence is exhaustively portrayed by a color image. The encoded data of multiple synchronized sensory devices is concatenated for comprehensively capturing high-level spatiotemporal correlations, not only within a sensor but also between different sensors. The activity recognition deep model is learned by end-to-end fine-tuning a pre-trained Inception-v3 network. In experiments, with the outstanding results on two common multi-sensors datasets of daily life activities, the proposed method defeats other existing approaches in terms of recognition accuracy.

II. METHODOLOGY

The overall workflow of our proposed method for daily life activity recognition based on wearable sensors is presented in Fig. 1, wherein the pre-processing stage includes segmentation for partitioning sequential data into fixed windows, data encoding for converting tri-axial inertial data to image, and image-formed representation for visualizing a segmented sample by a color image. The recognition model learning is done by end-to-end fine-tuning a pre-trained CNNs model. It should be noticed that feature engineering processes are automatically performed during training/fine-tuning a deep network.

A. Data Encoding: From Inertial Value To Pixel

Given a segmented inertial sequence $S = \{s_1, \ldots, s_N\}$, consisted of N tri-axial samples (from either accelerometer or gyroscope), wherein each sample is organized by three values corresponding to three axes of X, Y, and Z. Due to the differentiation of measurement unit, the inertial data has to be normalized before converting to chromatic RGB value. It should be noticed that the normalization process is performed for accelerometer and gyroscope separately by a general equation as follows

$$s' = \frac{s - \min(s)}{\max(s) - \min(s)}. \tag{1}$$

The values of normalized inertial data are distributed in the range $[0, 1]$. Each sample in the proposed scheme is encoded to be a color pixel, wherein three axis values are signified for three color channels of red, green, and blue, respectively. This encoding process is performed by a general intensity conversion function as follow

$$p = (g_{\max} - g_{\min}) \times s', \tag{2}$$

where p refers to as a color pixel and $[g_{\min}, g_{\max}]$ is the range of gray-scale value (usually $g_{\min} = 0$ and $g_{\min} = 255$ for the full-scale conversion of 24-bits RGB color image). The conversion function is used for both accelerometer and gyroscope data after normalizing since they are represented in a same unit. For example, from an original acceleration sample $s = (a_X, a_Y, a_Z)$, the normalized one s' is obtained by (1). According to the intensity conversion function, the corresponding color pixel p is formed as $p = (g_R \leftarrow a_X, g_G \leftarrow a_Y, g_B \leftarrow a_Z)$ from the sample s (see an illustration in Fig. 2a). At this point, a sequence S collected from an inertial sensor is completely portrayed by the array of color pixels (see an illustration in Fig. 2b), denoted $P = \{p_1, p_2, p_3, \ldots, p_N\}$.

B. Image-formed Representation

Corresponding to the segmented inertial sequence S of a single sensor, the color image I with the size of $1 \times N$ is established from the array of color pixels P, as follows

$$I = \begin{bmatrix} p_1 & p_2 & \ldots & p_N \end{bmatrix}. \tag{3}$$

In the case of multiple sensors, an activity image I with the size of $k \times N$ is constructed by concatenating several pixel arrays as follows

$$I = \begin{bmatrix} P_{S_1} \\ P_{S_2} \\ \vdots \\ P_{S_k} \end{bmatrix} = \begin{bmatrix} p_1^{S_1} & p_2^{S_1} & \ldots & p_N^{S_1} \\ p_1^{S_2} & p_2^{S_2} & \ldots & p_N^{S_2} \\ \vdots & \vdots & \ddots & \vdots \\ p_1^{S_k} & p_2^{S_k} & \ldots & p_N^{S_k} \end{bmatrix}, \tag{4}$$

where k refers to as the number of sensors (e.g., the number of accelerometer and gyroscope). It should be noticed that all sensors have to be configured at a same sampling frequency.

C. Recognition Model Learning: Fine-Tuning CNNs

Recently, deep CNNs have achieved state-of-the-art performance in terms of accuracy for many visual classification tasks. Inception-v3, inspired by GoogleNet (a.k.a., Inception-v1), is much enhanced with convolution factorization, regularization of auxiliary classifier, and efficient grid size reduction. In particular, Inception-v3 mainly structured by three types of inception module as follows: Inception-A with a sequence of 3×3 convolutional layers for shrinking the number of parameters, Inception-B with several asymmetric

convolutional layers 1×7 and 7×1 for reducing computational cost, and Inception-C with multiple filter banks for conducting high dimensional sparse representation. Instead of training a network from scratch with randomly initialized weights, a pre-trained CNN model with transfer learning technique is mostly adopted when dealing with another problems. Therefore, in this research, we fine-tune end-to-end a pre-trained Inception-v3 network model with a small learning rate coupled with dropout for overfitting prevention, besides modifying the last fully connected layer to fit the number of activities. The goal of fine-tuning is the minimization of the cross-entropy loss function as follows

$$\zeta(\mathbf{y}, \widehat{\mathbf{y}}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{C} \mathbf{y}_{ij} \log \widehat{\mathbf{y}}_{ij}, \tag{5}$$

where $\mathbf{y}$ and $\widehat{\mathbf{y}}$ refers to as the ground-truth label and the predicted label, respectively; n refers to as the number of samples in a batch and C is the number of activity classes.

III. Experiments

In this section, the proposed method is evaluated on two well-known multi-sensor datasets of daily life activities: Daily and Sport Activities [6] and Daily Life Activities (DaLiAc) [8]. The method is further compared with other existing HAR approaches.

A. Datasets

Daily and Sport Activities: This dataset is collected by using five body-worn miniature inertial sensor units (including an accelerometer, a gyroscope, and a magnetometer) from eight participants. There are totally 9120 5-seconds segmented sequences, recorded at 25 Hz sampling frequency, which represent for 19 activities.

Daily Life Activities (DaLiAc): This dataset, accumulated by four shimmers including accelerometers and gyroscopes, contains the inertial data of 19 subjects, in which each subject continuously performs 13 activities. All sensors are synchronized for data collection with 200 Hz sampling rate.

B. Experimental Setup

The method is benchmarked on two datasets following the leave-one-subject-out cross-validation (LOSOCV) [6], where one subject is for testing and remaining subjects are for training. The average recognition accuracy of m times (where m refer to as the number of subjects) is given. Regarding to learning recognition model, we fine-tune a pre-trained Inception network in 16 epochs using the stochastic gradient descent with momentum (SGDM) optimizer, the mini-batch size of 64, the initial learning rate of 0.001 and dropped 90% after 8 epochs.

C. Results & Discussion

The confusion matrices of the proposed activity recognition method are presented in Fig. 3 for the Daily and Sport Activities dataset and Fig. 4 for the Daily Life Activities dataset. It is worth that, in Fig. 3, several sitting samples are misclassified to the standing and lying classes. Some other confusions are occurred with activities in an elevator and exercises on a treadmill. Meanwhile, the experimental result in Fig. 4 indicates that two bicycling activities in DaLiAc are much confused each other for classification.

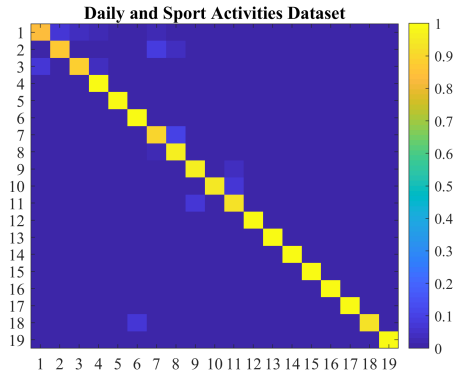


Fig. 3: The confusion matrix of the proposed method on Daily and Sport Activities Dataset withe overall accuracy of 95.8%, where the class 1 to 19 standing for sitting, standing, lying on back, lying on right side, ascending stairs, descending stairs, standing in an elevator, moving around in an elevator, walking in a parking lot, walking on a treadmill in flat, walking on a treadmill in 15^0 inclined position, running on a treadmill, exercising on a stepper, exercising on a cross trainer, cycling in horizontal position, cycling in vertical position, rowing, jumping, and playing basketball.

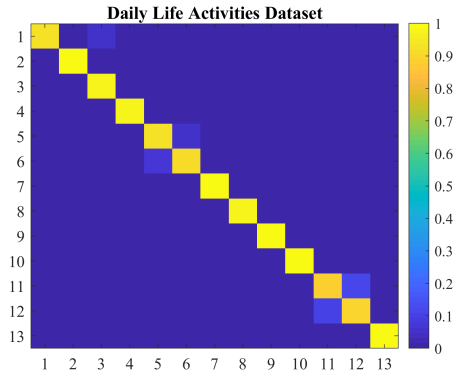


Fig. 4: The confusion matrix of the proposed method on Daily Life Activities Dataset withe overall accuracy of 95.7%, where the class 1 to 13 referring to as sitting, lying, standing, washing dishes, vacuuming, sweeping, walking outside, ascending stairs, descending stairs, treadmill running, bicycling (50 watt), bicycling (100 watt), and rope jumping.

The proposed method, denoted AcImgEncoding + CNNs, is further compared with several state-of-the-art HAR approaches, in terms of recognition accuracy, shown in Table I and II of two activity datasets, respectively. For the Daily and Sport Activities dataset, AcImgEncoding + CNNs significantly outperforms traditional methods [6] (e.g., the combination of statistical features and classical classification

TABLE I: Method Comparison on Daily and Sport Activities dataset

Method	Accuracy (%)
Statistical Feature + BDM [6]	75.8
Statistical Feature + DTW [6]	85.2
Statistical Feature + LSM [6]	85.3
Statistical Feature + *k*-NN [6]	86.9
Statistical Feature + SVM [6]	87.6
FTA + *k*-NN [7]	85.5
pFTA-Random + *k*-NN [7]	86.9
Ensemble [7]	87.1
Temporal Aggregate [7]	88.3
LSTM [7]	88.9
dRNN [7]	89.7
pFTA-Learn + *k*-NN [7]	90.2
Proposed AcImgEncoding + CNNs	**95.8**

TABLE II: Method Comparison on Daily Life Activities dataset

Method	Accuracy (%)
Joint Sensors [9]	82.1
Posterior Probability [9]	86.8
Hierarchical Classifier [8]	89.6
Independent Bayesian Classifier [9]	92.6
Markov Soft Output Classifier (MSCC) [9]	95.4
Proposed AcImgEncoding + CNNs	**95.7**

techniques) and modern methods [7] (e.g., advanced feature extraction algorithms and deep learning). For example, AcImgEncoding + CNNs is more accurate than pFTA-Learn + *k*-NN [7] by 5.6%. By encoding and representing the whole information of an activity sample to an image, high-level discriminative features, including the spatial correlations between multiple sensors and the temporal correlations between tri-axial values within a sensor, are automatically extracted over convolution and pooling operations at multi-levels of feature representation. For the Daily Life Activities dataset, AcImgEncoding + CNNs reaches the highest accuracy, compared with others that have exploited hierarchical classification architecture [8] or advanced classification models [9], but the improvement is minor. However, the simplicity of fine-tuning a single CNNs model is an obvious benefit for the practical implementation of healthcare applications over cloud computing.

IV. CONCLUSIONS

In this paper, we introduced a novel technique of converting inertial data to color image for wearable sensor based daily life activity recognition. This data encoding technique coupled with fine-tuning a pre-trained network model allows of efficiently learning a recognition model from activity images. By converting tri-axial samples, collected from inertial sensors like accelerometer and gyroscope, to color pixels (three axes values corresponding to three color channels), an activity image is generated by concatenating multiple pixel arrays. Based on experimental results, the proposed method is better than several state-of-the-art HAR approaches when conducting the overall accuracy of higher than 95% for the Daily and Sport Activities and the Daily Life Activities datasets.

ACKNOWLEDGMENT

This work was supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2018R1A6A1A03024003).

REFERENCES

[1] O. Banos et al., "Mining human behavior for health promotion," *in Proc. 2015 37th Ann. Int. Conf. IEEE Eng. Med. Biol. Soc.y (EMBC)*, Milan, 2015, pp. 5062-5065.

[2] N. A. Tu, T. Huynh-The, K. U. Khan and Y. Lee, "ML-HDP: A Hierarchical Bayesian Nonparametric Model for Recognizing Human Actions in Video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 800-814, March 2019.

[3] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, "Deep learning for sensor-based activity recognition: A Survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3-11, March 2019.

[4] R. Saeedi, K. Sasani, S. Norgaard and A. H. Gebremedhin, "Wavelet-Based Sit-To-Stand Detection and Assessment of Fall Risk in Older People Using a Wearable Pendant Device," *in Proc. 2018 40th Ann. Int. Conf. IEEE Eng. Med. Biol. Soc.y (EMBC)*, Honolulu, HI, 2018, pp. 1193-1196.

[5] A. Ejupi, M. Brodie, S. R. Lord, J. Annegarn, S. J. Redmond and K. Delbaere, "An improved algorithm for human activity recognition using wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1602-1607, July 2017.

[6] K. Altun, B. Barshan, O. Tuncel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors," *Pattern Recognit.*,vol. 43, no. 10, pp. 3605-3620, 2010.

[7] J. Ye, G. Qi, N. Zhuang, H. Hu and K. A. Hua, "Learning Compact Features for Human Activity Recognition via Probabilistic First-Take-All," *IEEE Trans. Pattern Anal. Mach. Intell. (Early Access)*, 2018, doi: 10.1109/TPAMI.2018.2874455.

[8] H. Leutheuser, D. Schuldhaus, B. M. Eskofier, "Hierarchical, Multi-Sensor Based Classification of Daily Life Activities: Comparison with State-of-the-Art Algorithms Using a Benchmark Dataset," *PLoS One*, vol. 8, no. 10, 2013.

[9] A. Nazabal, P. Garcia-Moreno, A. Artes-Rodriguez and Z. Ghahramani, "Human Activity Recognition by Combining a Small Number of Classifiers," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 5, pp. 1342-1351, Sept. 2016.

[10] S. Liu, R. X. Gao, D. John, J. W. Staudenmayer and P. S. Freedson, "Multisensor Data Fusion for Physical Activity Assessment," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 687-696, March 2012.

[11] A. Murad and J. Y. Pyun, "Deep Recurrent Neural Networks for Human Activity Recognition," *Sensors*, vol. 17, no. 11, 2017.

[12] J. He, Q. Zhang, L. Wang and L. Pei, "Weakly Supervised Human Activity Recognition from Wearable Sensors by Recurrent Attention Learning," *in IEEE Sensors J. (Early Access)*, 2018, doi:0.1109/JSEN.2018.2885796.

[13] M. Inoue, S. Inoue and T. Nishida, "Deep recurrent neural network for mobile human activity recognition with high throughput," *Artif. Life Robot.*, vol. 23, no. 2, pp. 173-185, 2018.

[14] F. J. Ordonez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, 2016.

[15] W. Chen, C. A. Betancourt Baca and C. Tou, "LSTM-RNNs combined with scene information for human activity recognition," *in Proc. 19th IEEE Int. Conf. e-Health Netw. Appl. Serv. (Healthcom)*, Dalian, 2017, pp. 1-6.

[16] T. Zebin, M. Sperrin, N. Peek and A. J. Casson, "Human activity recognition from inertial sensor time-series using batch normalized deep LSTM recurrent networks," *in Proc. 2018 40th Ann. Int. Conf. IEEE Eng. Med. Biol. Soc.y (EMBC)*, Honolulu, HI, 2018, pp. 1-4.

[17] M. Zeng et al., "Convolutional Neural Networks for human activity recognition using mobile sensors," *in Proc. 6th Int. Conf. Mobile Comput. Appl. Serv.*, Austin, TX, 2014, pp. 197-205.

[18] T. Hur et al., "Iss2Image: A Novel Signal-Encoding Technique for CNN-Based Human Activity Recognition," *Sensors*, vol. 2018, no. 11, 2018.