International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:14, No:7, 2020

Vision-Based Daily Routine Recognition for Healthcare with Transfer Learning

Bruce X. B. Yu, Yan Liu, Keith C. C. Chan

***Abstract*—**We propose to record Activities of Daily Living (ADLs) of elderly people using a vision-based system so as to provide better assistive and personalization technologies. Current ADL-related research is based on data collected with help from non-elderly subjects in laboratory environments and the activities performed are pre-determined for the sole purpose of data collection. To obtain more realistic datasets for the application, we recorded ADLs for the elderly with data collected from real-world environment involving real elderly subjects. Motivated by the need to collect data for more effective research related to elderly care, we chose to collect data in the room of an elderly person. Specifically, we installed Kinect, a vision-based sensor on the ceiling, to capture the activities that the elderly subject performs in the morning every day. Based on the data, we identified 12 morning activities that the elderly person performs daily. To recognize these activities, we created a HARELCARE framework to investigate into the effectiveness of existing Human Activity Recognition (HAR) algorithms and propose the use of a transfer learning algorithm for HAR. We compared the performance, in terms of accuracy, and training progress. Although the collected dataset is relatively small, the proposed algorithm has a good potential to be applied to all daily routine activities for healthcare purposes such as evidence-based diagnosis and treatment.

***Keywords*—**Daily activity recognition, healthcare, IoT sensors, transfer learning.

I. Introduction

As the world population is aging, how we can provide the elderlies with the care and support that they need to live independently becomes increasingly important. The proportion of elderly people with age of 65 or over is projected to reach 1 billion by 2030, 1.5 billion by 2050 and 2.9 billion by 2100. In 1990, 54 million of the elderly population were aged 80 or over. This number nearly tripled to 143 million in 2019. Globally, the number of people over 80 years of age is projected to nearly triple again to 426 million by 2050 and this number is projected to be 881 million by 2100 [1]. In other words, the number of people over 80 years of age is predicted to grow much faster than those above 65. The risk of elderly people suffering from various Noncommunicable Diseases (NCDs) will be much higher than those that are younger. In fact, NCDs are collectively responsible for almost 70% of all deaths worldwide [2]. In addition to all these diseases, it is reported [2] that Alzheimer's diseases or other degenerative brain diseases among elderly people have been increasing more significantly when compared with other NCDs. Amidst such diseases, it is important that the elderly can maintain the ability to live independently. To alleviate the healthcare burden, it is important to ensure the physical and mental well-being of the elderly people are looked after. To do so, clinicians utilize various metrics observed from basic ADLs of the elderly as an important indicator of the level of autonomy they enjoy [3]. For example, if the level of ADLs is considered sufficient, it can indicate the slowing down of mental illness. If level of ADLs is considered insufficient, this could suggest that elderly people increase physical activity as an effective strategy to maintain independence [4]. If information about activities in the daily routine could be automatically collected, it could serve as crucial reference for the prevention of NCDs and for prescription of behavioral therapies.

Research in the past mainly focused on developing solutions to determine abnormality in some specific activities such as walking imbalance [5], falling [6], sitting down and standing up [7], etc. There has been some effort to investigate into the automatic recognition of ADLs such as cooking [8], bathing [9], medication intake [10], etc. However, since the development of some physical and cognitive disfunctions usually takes years to become noticeable, it could be too late by the time they are discovered for any effective actions to be taken. In other words, abnormal activities like falling, and walking balance problems should be prevented instead of being detected. When symptoms of NCDs become noticeable at their late stages, such abnormality detection methods could be useful for diagnosis through analyzing some specific activities that provide abnormality information. However, they might not be sufficient for elderly patients' lifestyle management which is essential for maintaining their independence for taking effective behavioral therapies and other treatments that could prevent the development of various NCDs.

Given that the recognition of ADLs could facilitate NCDs early and this could mean facilitating the independent living of the elderly so that their health and rehabilitation process can be monitored, and possible diseases detected early [11], we develop tools to automatically collect ADLs data and detect various activities all through the day. If they can be accurately detected, the daily routines could be much more easily detected and monitored to keep track of health conditions. Based on the data collected, even minor changes in ADLs can be detected and analyzed so as to ensure that elderlies can live a healthy lifestyle and to discover hidden diseases or physical dysfunctions. We propose a solution here that can be deployed in a smart home that integrates healthcare features like health

Bruce X. B. Yu is with the Department of computing, The Hong Kong Polytechnic University, Hong Kong (corresponding author, phone: 852-9060-2458; e-mail: bruce.xb.yu@connect.polyu.hk).

Yan Liu and Keith C. C. Chan are with the Department of computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: yan.liu@polyu.edu.hk, keith.chan@polyu.edu.hk).

monitoring, daily reminding, and disease prevention [12].

Existing technologies that enable home automation have received much attention recently as they can improve comfort for seniors when they live at home [13]. However, current smart homes lack activity recognition features and this could hinder the health monitoring. To overcome the problem, we therefore propose IoT sensor based HAR as the essential module for future smart home. This idea is captured in Fig. 1 which illustrates a general procedure that enables a home environment to have healthcare features. The proposed smart home environment is equipped with IoT sensors for data collection so as to facilitate HAR data. Such data can be post-processed to obtain useful information including semantic meaning and interpretation for healthcare.

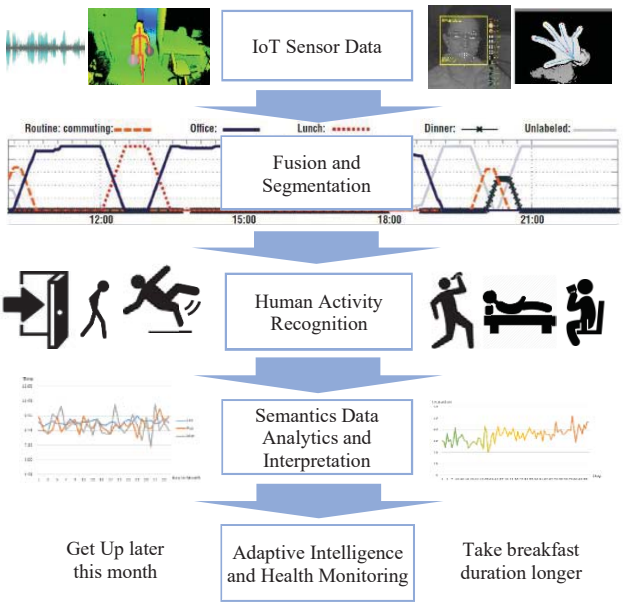

Fig. 1 Procedure of using IoT sensor based HAR to enable smart home environment with healthcare ability

Sensor based HAR has received much recent attention. However, there has not been much investigation regarding whether or not existing HAR methods could be applied to real-world environments in healthcare. Motivated by developing ADLs logging tool for elderly healthcare, we present here results of our effort to investigate feasible sensor deployment in the home environment and conduct a field study with an elderly person using video and infrared sensor as provided by Kinect v2. A real-world data set was collected on the daily routines performed by the elderly person and the dataset contained 12 regular ADLs in the morning. Based on our investigation, we developed an effective HAR algorithm by using transfer learning. In this paper, we also propose a framework that includes various HAR methods for implementation in an ambient elderly care environment. We conducted several experiments by following the HAR framework and achieved high activity recognition accuracy. We believe that the proposed method has good potential as it is simple yet effective in dealing with ADLs in ad-hoc healthcare environments that require monitoring physical independence ability parameters for evidence-based diagnosis and treatment.

II. Related Work

The body of literature related to the recognition of ADLs for healthcare is rare. We roughly classify related jobs according to the HAR sensors that are taken to tackle the problem. HAR methods can be classified according to the sensors and their arrangements. In this section, we review the type of IoT sensors and discuss how they are related to HAR approaches that are taken. We also consider the arrangement of the sensors and how they be utilized for the ADLs' recognition.

A. HAR Sensors

The HAR approaches taken to tackle the ADLs recognition problem can be depending on the types of sensors used. In [14], these sensors are classified roughly into two categories as ambient sensors and wearable sensors. Ambient sensors refer to sensors connected as a wireless mesh or dense network that monitors the whole indoor environment including human subjects. Wearable sensors refer to sensors that are attached to clothing or on the body, or even implanted under the skin. In [15], sensors are classified into three categories: vision, ambient and wearable. Following this categorization, we classify HAR literatures into three categories of methods [16]-[19]: vision-based, ambient sensor-based and wearable sensor-based methods. The vision-based approaches are the mainstream. Vision-based solutions usually make use of different vision features like silhouette, RGB, depth, and skeleton etc. for HAR. Ambient sensor-based approaches could have various ambient sensors like RFID tags, state change sensors and even Wi-Fi based sensors to choose from. As for the third category of approaches, they cover inertial sensors like accelerometer and gyroscope. Some recent works [22, 23] make use of different sensor modalities and develop sensor fusion methods to examine their mutual supplemental advantages so as to gather the most information for the most accurate HAR.

B. Sensor Arrangements

Recent work in the development of HAR methods depends on the arrangement of different sensors. For ambient sensors, some RFID technologies, such as that reported in [20], require that RFID tags are installed on the entire floor of a user's living environment for the purpose of detecting whether or not a person is near the bed. They also require that RFID antenna are embedded in the bed cloth. The main disadvantage is that, due to the interference of RFID signals when two objects are close together, detection accuracy will be low due to the high noise level. Another type of ambient sensor is state-change sensors that require to be installed in all locations for deployment. However, despite the wide adoption of state-change sensors, this approach could only do some coarse-grained HAR. In wireless communications, the channel state information (CSI) is known as channel properties of a communication link. Wi-Fi CSI is a useful approach that can be adopted for HAR to reduce the requirement for the number of sensors [21]. It also has the

TABLE I
DIFFERENT SENSOR ARRANGEMENTS USED IN EXISTING PUBLIC HAR DATASETS

Dataset	Year	Sensor	Sensor Type	NS	NA	Activity
MSRDailyActivity3D [26]	2012	Kinect v1	Vision	10	16	Actions
PKU-MMD [27]	2017	Kinect v2	Vision	66	51	Actions
NTU RGB+ D 120 [28]	2019	Kinect v2	Vision	32	120	Actions
Kasteren Dataset [29]	2008	State change	Ambient	1	8	Daily activity
Freiburg Dataset [30]	2014	Audio	Ambient	--	22	Daily activity
Smart Carpet Dataset [20]	2016	RFID	Ambient	13	2	Fall detection
WiAR Dataset [31]	2018	Wi-Fi	Ambient	10	16	Gestures, activities
PAMAP2 [32]	2012	3 3-DOF IMUs	Wearable	9	18	Daily activities
Opportunity dataset [22]	2011	IMUs, 72 sensors of 10 modalities	Mixed	12	21	Morning activities
Berkeley MHAD [23]	2013	Mocap, Kinect v1, camera, acc, audio	Mixed	12	11	Actions

advantage of cross wall sensing ability, but this approach lacks theoretical foundation that elaborates its accuracy and capability for multi-user activity recognition. Wi-Fi CSI based HAR requires strict sensor positioning, rendering it hard to install and adapt to environmental changes. Wi-Fi CSI is only used at its early research stage and it lacks comparison with other sensors in terms of its measurement accuracy.

Wearable devices could be an appropriate choice for activity recognition. Given that each sensor modality has its own limitations, there has been some effort to fuse vision and inertial sensor data to improve the HAR accuracy [22, 23]. In [24], a review of previous works that use both depth cameras and inertial sensors to collect multimodal data has been presented. It provides a summary of the similarities in the features that are utilized for such fusion approaches. However, the inertial modality does not provide any contextual information for fine-grained (e.g. human-object interaction) HAR tasks. Besides, due to its intrinsic battery limitation, this approach is considered too intrusive as batteries need to be replaced to allow devices being worn for long-term monitoring.

It is still unclear whether or not adding extra modalities can improve the accuracy of multimodal HAR methods. Based on the Opportunity dataset, the multimodal fusion analysis of [25] reveals that the more data channels there are for its proposed Deep Learning (DL) model named DeepConvLSTM, the better HAR performance can be. For example, starting from a F_1 score of 69% that is given by using only the accelerometer data modality of the Opportunity dataset, the average performance improves 15% by fusing accelerometers and gyroscopes and 20% when fuses accelerometers, gyroscopes and magnetic channels. However, the use of different data modality combinations in experiments based on Berkeley Multimodal Human Activity Database (MHAD), the improvement on the performance is very limited when adding more data modalities (from around 98% to 100%) [23]. It also concludes that adding extra modality may even lower the HAR accuracy, which renders the extra modality misleading for the recognition process. Besides, the increased problem complexity and difficulties for deployment make multimodal HAR hard to be popularized among end users and other stakeholders.

To compare the capability of different sensors, we collected some representative public datasets that use various sensor arrangements as listed in Table I. It is noteworthy that the NTU RGB+D 120 [28] includes relatively more complex activities comparing with the other datasets. This dataset is collected based on vision sensors. Based on the analysis of existing sensor arrangements in Table I, it appears that vision sensors are relatively more reliable for HAR among all other sensors that have been adopted when it comes to the number of subjects (NS) involved in their datasets, and the number of activity classes (NA) that they try to recognize. However, with the ambition to simultaneously recognize activities with different activity complexities like activity resolution, high- and low-level activities, and human-object interactions, a dataset, NTU RGB+D 120 [28], containing a larger number of activities has been made available for testing. So far, this goal has not been achieved well as performance with the dataset suffers from relatively low accuracy. The most accurate recognition rate achieved is around 65% [28]. Also, if ADLs recognition for NCDs are to be tackled, the NTU RGB+D 120 does not need to be used in all when developing models for the task. Some fine-grained activities in the dataset like “make ok sign”, “counting money” and activities labelled as “grab other person’s stuff”, “put on bag/backpack”, and “put on jacket” might be irrelevant to ADLs or for inferring symptoms of NCDs. Besides, with the increasing deployment complexity, large datasets might be unfeasible when developing models for real-world healthcare applications.

III. COLLECTING AN ADLS DATASET FOR RESEARCH IN ELDERLY CARE

In this section, we introduce the data collection method of our ADLs dataset and its data structure. In [33], a range of ADLs are defined so that ADLs can be divided into domestic activities, or Instrumental ADLs (IADLs), and personal selfcare activities, or Basic ADLs (BADLs). IADLs are activities that are not essential but are needed for an individual to be able to live independently in a community. They include, for example, preparing food, housekeeping, shopping, managing money, taking medication, using telephone and transportation. BADLs refer to selfcare activities, such as eating, drinking, dressing, walking, bathing, and using toilet. Collecting these ADLs is the goal of life-logging systems like the one developed by Ahmad et al. [34], which uses depth video sensor to detect activities like cooking, watching TV, exercise, hand clapping, walking and cleaning. However, most life-logging systems, such as that described in [34], does not consider these characteristics of ADLs. Instead, only data related to some coarse-grained activities are collected. In this paper, we presented how we

collected a set of ADLs of an elderly to study and recognize her morning routines like "lie down", "get up", "comb hair", "sweep the floor", etc. that are essential for inferring health related information like healthy lifestyle, hygiene, sleeping hour, physical efficiency, etc.

A. Sensor Selection

To decide what sensors to best collect data for ADLs recognition, we have investigated into the relative advantages and disadvantages of each sensor category. As shown in Table II, we considered installation complexity, intrusiveness, and battery life and concluded that ambient and wearable sensors have relatively more disadvantages and may not be the most suitable for home based daily routine recognition. Having considered all these different sensors, we decided to go with vision-based sensors. Of different vision-based sensors, we decided to use Kinect. Specifically, we have chosen the Kinect v2 sensor. The valid working distance range of Kinect v2 sensor is 0.5 to 4.5 meters, which is long enough to cover the whole room environment of an elderly person. Unlike other RGB vision sensors, the Kinect v2 sensor will not be affected by poor illumination condition at night. Kinect provides skeleton retrieval technology, which reduces the video data volume and complexity of vision-based solution. The issue of privacy is a stereotype that can also be tackled with such technology as we only use and kept skeleton data that includes only 25 joints as shown in Fig. 2 and such data is not sufficient for the identity of a person to be revealed. To avoid the occlusion problem, we installed the Kinect sensor on the ceiling to cover as much area as possible without occlusion. This way, of course, means that the algorithms used for HAR has to be specially designed to allow ADLs to be recognized accurately.

TABLE II
AVAILABLE SENSORS FOR HAR

Sensor type	Video sensor	Ambient sensor	Wearable sensor
Sensor	Kinect v1/v2 MoCap Intel RealSense Stereo cameras Single cameras	Pressure/force Passive Infrared RFID Wi-Fi Microphone Ultrasonic	Inertial Measurement Units (IMUs) Biosensors GPS EEG ECG
Sensor/data	Depth RGB Motion/skeleton Infrared	Light Sound Motion Door Vibration Pressure	Body temperature Heart rate Accelerometer Gyroscope ECG/EEG Steps
Advantage	Nonintrusive	Nonintrusive	Location unlimited
Disadvantage	Occlusion/view point limited Light condition Pervasiveness Computational cost Privacy	Location limited Installation complexity Maintenance	Intrusive/obtrusive Acceptance of subject Battery life

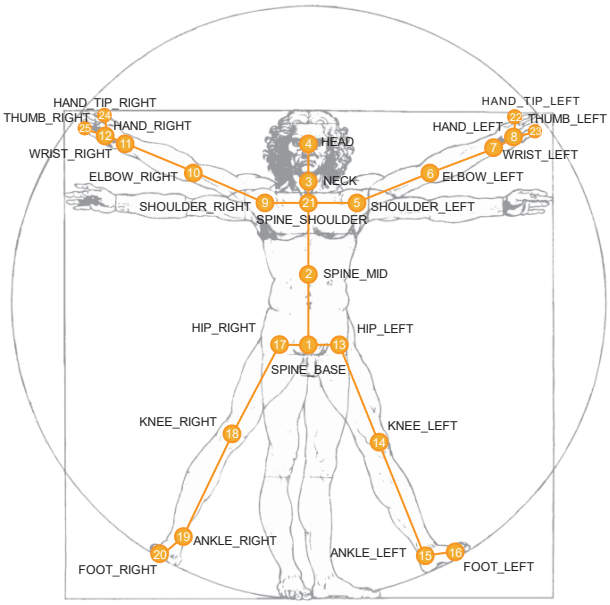

Fig. 2 Skeleton joints of Kinect v2 sensor

B. Sensor Installation

When other benchmarking datasets such as PKU-MMD [27] or NTU RGB+D [28] were collected, video sensors were usually mounted in front of the subjects. The problem with this is that the subjects could easily be occluded in real environments. For our case, the video sensor is mounted on the ceiling so as to cover the whole monitoring environment as much as possible. The use of fewer video sensors also has the benefit that labeling time can be reduced. With this sensor arrangement, we collected a dataset that is small yet sufficient for developing skeleton based HAR models that can be used for home healthcare. Fig. 3 shows two examples of ADLs that we collected data for. We provide here 3 sampled RGB frames captured by the Kinect v2 sensor.


Fig. 3 Sample frames of "Eat with chopsticks" and "sweep the floor"

C. Collecting ADLs

Publicly available datasets, such as the NTU RGB+D 120 dataset, is expanded from the NTU RGB+D 60 dataset by adding more fine-grained activities like hand or finger motions and object-related individual actions. It also adds more

challenging activities that share some similarities like similar body motions, similar objects, and similar gestures. These public datasets are usually collected for developing new HAR algorithms that can improve over existing algorithms by being able to recognize greater number of activities, faster detection speed with higher accuracy. However, not all these activities are relevant to ADLs. For the purpose of our applications, we examine the characteristics of ADLs of an elderly person living independently and collected a dataset from the subject that includes activities that the subject performs in the morning (see Table III). The dataset is different from existing benchmark datasets in three aspects. First, it is collected in a real environment and the activities are performed naturally. Unlike other datasets, the subject does not act for the purpose of data collection. Second, it is collected for recognizing daily routines that involve ADLs with proper granularity. Third, activities are collected over a period creating a dataset of size that is large enough for training a recognition model.

TABLE III
ADLS IN THE COLLECTED MORNING ROUTINE DATASET

Activity Label	Activity Name	ADLs Type	Times
1	lie down	BADLs	9
2	get up	BADLs	9
3	comb hair	BADLs	11
4	pour water	BADLs	9
5	drink water	BADLs	9
6	eat with chopsticks	BADLs	10
7	eat with irons poon	BADLs	10
8	eat with pottery spoon	BADLs	12
9	tidy table	IADLs	16
10	wipe table	IADLs	9
11	sweep the floor	IADLs	19
12	wear shoes	BADLs	17

As discussed above, we used the Kinect v2 sensors to collect our dataset. For any one particular activity being monitored, using the Kinect v2, we record a sequence of skeleton body frames corresponding to the actions performed. Each skeleton body frame consists of 25 joints (see Fig. 2) which can be labelled as HEAD, NECK, …, FOOTLEFT, etc. For a set of joints in a body frame that is observed at time *t*, let us represent the set as $\boldsymbol{j}^t = (\boldsymbol{j}_1^t, \dots, \boldsymbol{j}_i^t, \dots, \boldsymbol{j}_{25}^t)$ where $\boldsymbol{j}_i^t$ is the 3-D cartesian coordinates of the position of joint i so that $\boldsymbol{j}_i^t = (j_{ix}^t, j_{iy}^t, j_{iz}^t)$ with j_{ix}^t, j_{iy}^t, and j_{iz}^t correspond to the values of the x-, y- and z-coordinates, respectively. An activity that begins at time $t = 1$ and ends at time T with body frames collected at regular intervals can, therefore, be represented as a time series of T skeleton frames, $\mathbf{J_i} = [\boldsymbol{j}^1, \boldsymbol{j}^2, \dots, \boldsymbol{j}^t, \dots, \boldsymbol{j}^T]$. With M training samples, we will have $\mathbf{J} = \{\mathbf{J}_1, \mathbf{J}_2, \dots, \mathbf{J}_i, \dots \mathbf{J}_M\}$ for training.

IV. HARELCARE: A FRAMEWORK FOR ELDERLY ACTIVITY RECOGNITION

In this section, we present a framework for the recognition of ADLs using a machine learning algorithm we propose. This framework covers all steps it takes to process raw data collected from the sensor selection to the recognition of ADLs. Under an analysis of the existing methods within the framework, a transfer learning algorithm method is introduced for ADLs recognition.

A. ADLs Recognition Framework

The raw data that we make use of for HAR are obtained from a Kinect sensor and can be represented, as discussed above, as $\mathbf{J} = \{\mathbf{J}_i \mid i = 1, \dots, M\}$. This set of raw data is therefore a set of multivariate, spatial-temporal data. Traditionally, algorithms like the DTW [35], HMM [36], and SVM [37] have been proposed for developing predictive models for HAR based on skeleton data. More recently, deep learning algorithms [38] have been used for this task. The relative merits of these algorithms depend on such factors as accuracy, processing speed, and ease-of-deployment and there is always a need for us to develop an algorithm that can perform better according to these factors.

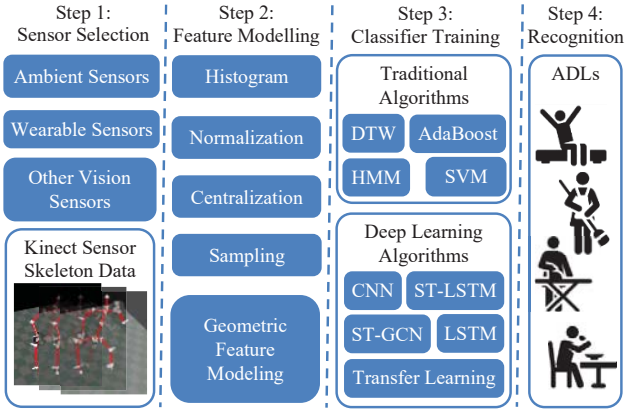

Fig. 4 HARELCARE: ADLs recognition framework

Towards this goal, we propose a 4 steps framework called HARELCARE (see Fig. 4) to better tackle the ADLs recognition. Under such a framework, we can develop a combination of different component algorithms to best address different problems and sub-problems. For example, we could make use of traditional feature modelling algorithms or more recent DL methods. For DL methods, feature modeling may or may not be necessary as DL methods may learn feature representation at the last fully connected layer.

In deciding what algorithms to develop for HAR, we note that algorithms with high processing speed could be easier to deploy but these algorithms usually perform with relatively lower accuracy. On the other than, algorithms that are computationally slow may perform better in terms of accuracy. For example, the use of the AdaBoost algorithm in the Visual Gesture Builder (VGB) tool [39] for the recognition of a single activity could be implemented with the proposed framework. Based on the use of a set of features selected from transformed raw data, the AdaBoost algorithm could be used effectively for single activity recognition with a confidence value ranging from 0 to 1. Using an extension of AdaBoost for the recognition of multiple activities could achieve an accuracy of 0.63 on the

MSRDailyActivity3D dataset [40].

For higher accuracy, many DL methods have been used for skeleton based HAR. Some make use of raw skeleton data as input and fed the data directly to DL models using such popular algorithms such as the Convolutional Neural Network (CNN) or Long Short-Term Memory (LSTM) algorithms for training [41, 42]. Some propose to use the context-aware LSTM algorithms [43] for training with the attempt to make the model focus on active skeleton joints that contribute more to the accuracy. There has also been some effort to remove noise in the skeleton data for view-invariant recognition using the approaches described in [44, 45]. In addition to these attempts, another potentially effective method for HAR is to use contextual information to improve HAR accuracy by modeling human-object interaction [46]. In addition to all these, there has recently been some effort to use multimodal approaches to HAR [47]. Instead of scaling up on the input data, spatial and temporal DL models like ST-LSTM [48] and ST-GCN [49] have been shown to be quite effective in handling the sparse skeleton data.

Even though these DL approaches are relatively more accurate, they require big datasets for training and it should be noted that it may not always be easy for big datasets to be collected. To avoid this problem, we propose a transfer learning method [50] that can be used for post-processing after the ST-GCN algorithm [49] is adopted.

B. Transfer Learning

Collecting less data could ease the deployment of activity recognition, however, DL models are usually face with overfitting when no sufficient data is available. Transfer learning that fine-tunes a pre-trained DL network weights from one task to another similar task have been proven helpful, which is a common strategy for transfer learning in the context of deep learning. [51] grouped transfer learning for HAR in three scenarios: inter-person, inter-device, and inter-ambiance. As ST-GCN [49] shows the potential for representing spatial and temporal features of skeleton data, we propose to use it as the backbone model for transfer learning. Precisely, we tune the weights of ST-GCN trained on the NTU-RGB+D dataset to our dataset. Our transfer learning method could be considered as inter-ambiance since we use different data collection environment with NTU-RGB+D. ST-GCN is basically a Graph Convolutional Network (GCN) designed to learn a representation of both spatial and temporal features from graph data. GCN is efficient to represent the sparse skeleton data, which is symbolized as $\boldsymbol{\vartheta}_t = \{\boldsymbol{\upsilon}_t, \boldsymbol{\varepsilon}_t\}$, where $\boldsymbol{\upsilon}_t$ denotes the skeleton joints and $\boldsymbol{\varepsilon}_t$ demotes the skeleton bones time t, respectively. A node υ_{ti} will have a neighbor set defined as $\mathbf{N}(\upsilon_{ti}) = \{\upsilon_{tj} | d(\upsilon_{ti}, \upsilon_{tj}) \leq D\}$, where D is to limit the length of $d(\upsilon_{ti}, \upsilon_{tj})$. Assuming that we have K subsets in every neighbor set $\mathbf{N}(\upsilon_{ti})$ of a node υ_{ti}, it will be indexed by a mapping $l_{ti}: \mathbf{N}(\upsilon_{ti}) \rightarrow \{0, \dots, K-1\}$. Then the convolutional operation of the graph could be calculated as

$$\mathrm{Y_{output}}\ (\boldsymbol{\upsilon}_{ti}) = \sum_{\boldsymbol{\upsilon}_{tj} \in \mathbf{N}(\upsilon_{ti})} \frac{1}{Z_{ti}(\upsilon_{tj})} \mathbf{X}(\upsilon_{tj}) \mathrm{W}(\mathrm{l}(\upsilon_{tj})) \quad (1)$$

where $\mathbf{X}(\upsilon_{tj})$ is the feature of $\boldsymbol{\upsilon}_{tj}$ that is equal to $(j_{jx}^t, j_{jy}^t, j_{jz}^t)$, $\mathrm{W}(l(\upsilon_{tj}))$ is a weight function $\mathrm{W}(\mathrm{v}_{ti}, \mathrm{v}_{tj}): \mathbf{N}(\upsilon_{ti}) \rightarrow \mathrm{R}^{\mathrm{c}}$ that could be implemented by indexing a tensor of (c, K) dimension. While $Z_{ti}(\upsilon_{tj}) = |\{\upsilon_{tk} | l_{ti}(\upsilon_{tk}) = l_{ti}(\upsilon_{tj})\}|$ is a normalization term that equals to the cardinality of the corresponding subset. With the specific partitioning strategy determined, Equation 1 could be implemented with adjacency matrix $\mathbf{A}$ as

$$\mathrm{Y_{output}} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \boldsymbol{\Lambda}_{\mathrm{k}}^{-\frac{1}{2}} \mathbf{A}_{\mathrm{k}} \boldsymbol{\Lambda}_{\mathrm{k}}^{-\frac{1}{2}} \mathbf{X} \mathrm{W}_{\mathrm{k}} \quad (2)$$

where $\boldsymbol{\Lambda}_{\boldsymbol{k}}^{\boldsymbol{ii}} = \sum_{\boldsymbol{j}} \mathbf{A}_{\boldsymbol{k}}^{\boldsymbol{ij}}$ is a degree matrix. Weiss et al. [52] performed a through survey for the transfer learning, which classifies transfer learning according to different categories as homogeneous transfer learning solutions, heterogeneous transfer learning solutions, and solutions addressing negative transfer that were further grouped to sub-categories. According to the categorization of [52], our method is homogeneous transfer learning as the feature space $\mathcal{X}_T$ of the target domain $\mathcal{D}_T$ (our dataset) has the same data structure with the feature space $\mathcal{X}_S$ of source domain $\mathcal{D}_S$ (the NTU-RGB+D dataset). We use the trained weights W_S from the feature space $\mathcal{X}_S$ to tune the weights W_T for $\mathcal{D}_T$. The transfer learning method is elaborated in Algorithm 1, which introduces the fine-tuning process.

Algorithm 1: Transfer Learning

Data: $\mathcal{X}_T$, the input data **J**
Output: Y_{output}, the output of the target domain
1. Load weights W_S that is trained by using $\mathcal{X}_S$
2. Modify output layers of ST-GCN to adapt the output $\mathrm{Y_{output}}$ of $\mathcal{D}_T$
3. Feed $\mathcal{X}_T$ to the modified model
4. **For** $i = 1$ **to** epoch M **do**
5. | **For** $j = 1$ **to** batch N **do**
6. | | Update W_T
7. | **End**
8. **End**
9. Use W_T to infer Y_{output}

V. Experiments

In this section, we explain how the experiments are set up to evaluate the performance of the proposed method.

A. Evaluation Measures

For performance evaluation, we perform cross-validation on all the above algorithms that we described to compare their *top-1* accuracy which is defined as

$$P = \frac{1}{N} \sum_{k=1}^{N} result_k = \begin{cases} 1 & if\ output_k^{top-1} = label_k \\ 0 & otherwise \end{cases}. \quad (3)$$

The *top-1* accuracy measure reflects the matching between the model prediction and the ground truth must be exactly the same. With the accuracy measure set as *top-1*, a confusion matrix, also known as error matrix, can be constructed as shown in Table IV. This matrix could be used to visualize the performance of a supervised classification algorithm with two or more classes [53]. To evaluate the models, we accumulate

five folds and normalize the entries in the confusion matrix for further comparison of the overall performances of different HAR algorithms.

TABLE IV
A SAMPLE OF CONFUSION MATRIX WITH 5 CLASSES

Actual Class \ Predicted Class	1	2	3	4	5
1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}
2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}
3	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}
4	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}
5	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}

B. Experiment Setting

For our experiments, we used k-fold cross-validation with k following the convention set to 5 [54]. During the training, all the other settings are the same except that we initialized the weights trained on NTU-RGB+D at epoch 80. The training procedure of our transfer learning method follows the Algorithm 1, where we replaced the final 2D convolutional layer (256, 60) of ST-GCN with a linear layer (256, 12). In the learning process, the epoch number in all experiments were set to 200. The leaning rate decay parameter of ST-GCN were empirically set at 40, and 100. All other hyper parameters are the same with the original setting. We set the interval of retrieving progressive training information to 10, which means it will record results of training mean average loss, testing mean average loss, and *top-1* accuracy with an interval of 10 epochs. All the training process and evaluation were run on a Supermicro GPU Server (model SYS-7048GR-TR) with 4 GTX 1080 Ti GPUs.

C. Experimental Results

Table V shows the *top-1* accuracy of both the ST-GCN model and our transfer learning algorithm. From the results we could observe that transfer learning achieved better *top-1* accuracy in every cross-validation fold. The average *top-1* accuracy of transfer learning is 91.64%, which is significantly higher than that of the ST-GCN (68.42%). It indicates the practical ability of transfer learning method for real-world healthcare applications when there is not enough training data.

TABLE V
TOP-1 ACCURACY ON ST-GCN AND TRANSFER LEARNING

CV Folds	ST-GCN	Transfer Learning
Fold 1	66.67%	83.33%
Fold 2	77.41%	93.10%
Fold 3	67.86%	92.86%
Fold 4	55.17%	93.10%
Fold 5	75.00%	95.83%
Average	68.42%	91.64%

Other than showing improvement of the *top-1* accuracy by using transfer learning, the confusion matrices of them are visualized to further investigate the improvement as shown in Table VI and Table VII, respectively. From Table VI, it is noted that ST-GCN could not performed well on some activities like "get up", "eat with chopsticks", and "eat with iron spoon". There is still some improvement space when tackle with a relatively small dataset that tends to be overfitted by DL models. In other words, if a dataset is too small, there is a need for local minimum to be avoided and for the model to be more effectively optimized. According to the results as presented in Table VII, the performance of the proposed transfer learning algorithm is close to optimum and there is little space for further improvement. Closer examination of the cases that the model failed to correctly recognize is mainly due to the big bias or unexpected noise like failure of skeleton detection by the Kinect v2 sensor.

TABLE VI
ACCUMULATED AND NORMALIZED CONFUSION MATRIX OF ST-GCN

True Label \ Predicted Label	1	2	3	4	5	6	7	8	9	10	11	12
1	0.78	0.22	0	0	0	0	0	0	0	0	0	0
2	0.67	0.33	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0.56	0	0	0	0	0	0	0	0.44
5	0	0	0	0.22	0.67	0	0	0	0	0	0	0.11
6	0	0	0	0.4	0	0.1	0.1	0.4	0	0	0	0
7	0	0	0	0.5	0	0	0.2	0.3	0	0	0	0
8	0	0	0	0.17	0	0.08	0.25	0.5	0	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0	0
10	0	0	0	0	0	0	0	0	0.11	0.89	0	0
11	0	0	0	0	0	0	0	0	0.37	0	0.63	0
12	0	0	0	0	0	0	0	0	0	0	0	1

TABLE VII
ACCUMULATED AND NORMALIZED CONFUSION MATRIX OF TRANSFER LEARNING

True Label \ Predicted Label	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	s0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0.11	0.67	0.11	0	0	0.11	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0
6	0	0	0	0	0	0.8	0	0.2	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0	0
8	0	0	0	0	0	0.08	0	0.92	0	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0	0
10	0	0	0	0.11	0	0	0	0	0	0.67	0.22	0
11	0	0	0	0	0	0	0	0.05	0	0	0.95	0
12	0	0	0	0.06	0	0	0	0.06	0	0	0	0.88

VI. CONCLUSION WITH DISCUSSION

This paper introduced an ADLs recognition framework called HARELCARE that could be used for daily routine collection for independent elderly people. We investigated the related works of HAR in terms of sensors and datasets for the

development of real-world ADLs recognition methods. Then we utilized the Kinect v2 sensor and adopted a transfer learning method onto the collected morning routine dataset. The activities in our dataset is based on the concern of NCDs prevention with a need to automatically collect ADLs. According to the experimental results, even small dataset could achieve great accuracy by using our transfer learning method. With such a promising result, our method has a great potential to be applied to healthcare application scenarios like habit perception, intervention performance evaluation, disease prediction, and adaptive (automatic) smart home.

Although we achieved high accuracy on the morning routine dataset, one current trend for HAR is using multimodal data for recognizing activities with higher resolutions. Weather adding extra modalities will improve the existing HAR tasks remains controversial. Hence, more research is required to investigate advanced HAR methods that provides more detailed information for evidence-based diagnosis and treatment. Besides, other sensors like RealSense could be used to get more fine-grained features like emotion, eye gaze, and facial expression, which will benefit real world scenarios like elderly home and houses of independent elderly people. Although medical datasets for various diseases are widely available, the lack of behavior datasets remains an issue for developing evidence-based therapies to prevent the elderly suffering from NCDs. As far as we know, there are very few field studies that last for a long activity monitoring period and conduct a long-term HAR based disease evaluation. With the high accuracy of our HAR method on the morning routine dataset in the real world environment, one of our future jobs is to collect and accumulate daily behavioral data by applying our method to homes of independent elderly people, and then analyze the behavioral data to infer symptoms related to NCDs.

REFERENCES

[1] U. DESA, "World Population Prospects 2019: Highlights," New York (US): United Nations Department for Economic and Social Affairs, 2019.
[2] J. Gill and M. J. Moore, "The State of aging & health in America 2013," 2013. Available: https://www.statista.com/statistics/207347/causes-of-death-among-us-adults-aged-65-by-ethnicity/.
[3] E. C. Nelson, T. Verhagen, and M. L. Noordzij, "Health empowerment through activity trackers: An empirical smart wristband study," Computers in human behavior, vol. 62, pp. 364-374, 2016.
[4] E. Tak, R. Kuiper, A. Chorus, and M. Hopman-Rock, "Prevention of onset and progression of basic ADL disability by physical activity in community dwelling older adults: a meta-analysis," Ageing research reviews, vol. 12, no. 1, pp. 329-338, 2013.
[5] M. Gabel, R. Gilad-Bachrach, E. Renshaw, and A. Schuster, "Full body gait analysis with Kinect," in Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, 2012: IEEE, pp. 1964-1967.
[6] Y. T. Liao, C.-L. Huang, and S.-C. Hsu, "Slip and fall event detection using Bayesian Belief Network," Pattern recognition, vol. 45, no. 1, pp. 24-32, 2012.
[7] A. Elkholy, M. Hussein, W. Gomaa, D. Damen, and E. Saba, "Efficient and Robust Skeleton-Based Quality Assessment and Abnormality Detection in Human Action Performance," IEEE journal of biomedical and health informatics, 2019.
[8] P. Lukowicz et al., "Recording a complex, multi modal activity data set for context recognition," in Architecture of Computing Systems (ARCS), 2010 23rd International Conference on, 2010: VDE, pp. 1-6.
[9] K. Chapron, P. Lapointe, K. Bouchard, and S. Gaboury, "Highly Accurate Bathroom Activity Recognition using Infrared Proximity Sensors," IEEE Journal of Biomedical and Health Informatics, 2019.
[10] C. Chen, N. Kehtarnavaz, and R. Jafari, "A medication adherence monitoring system for pill bottles based on a wearable inertial sensor," in Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, 2014: IEEE, pp. 4983-4986.
[11] G. Sprint, D. Cook, D. Weeks, J. Dahmen, and A. La Fleur, "Analyzing sensor-based time series data to track changes in physical activity during inpatient rehabilitation," Sensors, vol. 17, no. 10, p. 2219, 2017.
[12] Q. Ni, A. García Hernando, and I. de la Cruz, "The elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development," Sensors, vol. 15, no. 5, pp. 11312-11362, 2015.
[13] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects," Personal and Ubiquitous Computing, vol. 17, no. 1, pp. 127-144, 2013.
[14] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 6, pp. 790-808, 2012.
[15] F. Palumbo, J. Ullberg, A. Štimec, F. Furfari, L. Karlsson, and S. Coradeschi, "Sensor network infrastructure for a home care monitoring system," Sensors, vol. 14, no. 3, pp. 3833-3860, 2014.
[16] H. Alemdar and C. Ersoy, "Wireless sensor networks for healthcare: A survey," Computer networks, vol. 54, no. 15, pp. 2688-2710, 2010.
[17] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," Pattern Recognition Letters, vol. 48, pp. 70-80, 2014.
[18] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," IEEE sensors journal, vol. 15, no. 3, pp. 1321-1330, 2015.
[19] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," IEEE Communications Surveys and Tutorials, vol. 15, no. 3, pp. 1192-1209, 2013.
[20] A. Wickramasinghe, R. L. S. Torres, and D. C. Ranasinghe, "Recognition of falls using dense sensing in an ambient assisted living environment," Pervasive and mobile computing, vol. 34, pp. 14-24, 2017.
[21] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in Proceedings of the 21st annual international conference on mobile computing and networking, 2015: ACM, pp. 65-76.
[22] H. Sagha et al., "Benchmarking classification techniques using the Opportunity human activity dataset," in Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, 2011: IEEE, pp. 36-40.
[23] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in Applications of Computer Vision (WACV), 2013 IEEE Workshop on, 2013: IEEE, pp. 53-60.
[24] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," Multimedia Tools and Applications, vol. 76, no. 3, pp. 4405-4425, 2017.
[25] 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.
[26] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012: IEEE, pp. 1290-1297.
[27] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A Large Scale Benchmark for Skeleton-Based Human Action Understanding," in Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities, 2017: ACM, pp. 1-8.
[28] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan, and A. K. Chichung, "NTU RGB+ D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," IEEE transactions on pattern analysis and machine intelligence, 2019.
[29] T. Van Kasteren, A. Noulas, G. Englebienne, and B. Kröse, "Accurate activity recognition in a home setting," in Proceedings of the 10th international conference on Ubiquitous computing, 2008: ACM, pp. 1-9.
[30] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-markovian ensemble voting," in RO-MAN, 2012 IEEE, 2012: IEEE, pp. 509-514.
[31] L. Guo, L. Wang, J. Liu, W. Zhou, and B. Lu, "HuAc: Human Activity Recognition Using Crowdsourced WiFi Signals and Skeleton Data," Wireless Communications and Mobile Computing, vol. 2018, 2018.
[32] A. Reiss and D. Stricker, "Creating and benchmarking a new dataset for physical activity monitoring," in Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments, 2012: ACM, p. 40.

[33] S. Katz, "Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living," Journal of the American Geriatrics Society, vol. 31, no. 12, pp. 721-727, 1983.
[34] A. Jalal, S. Kamal, and D. Kim, "A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments," Sensors, vol. 14, no. 7, pp. 11735-11759, 2014.
[35] D. M. Gavrila and L. S. Davis, "Towards 3-d model-based tracking and recognition of human movement: a multi-view approach," in International workshop on automatic face-and gesture-recognition, 1995: Citeseer, pp. 272-277.
[36] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in Proceedings. Fourth IEEE International Conference on Multimodal Interfaces, 2002: IEEE, pp. 3-8.
[37] R. Lublinerman, N. Ozay, D. Zarpalas, and O. Camps, "Activity recognition from silhouettes using linear systems and model (in) validation techniques," in 18th International Conference on Pattern Recognition (ICPR'06), 2006, vol. 1: IEEE, pp. 347-350.
[38] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep Learning for Sensor-based Activity Recognition: A Survey," arXiv preprint arXiv:1707.03502, 2017.
[39] "Kinect tools and resources." Available: https://developer.microsoft.com/en-us/windows/kinect/tools.
[40] F. Lv and R. Nevatia, "Recognition and segmentation of 3-d human action using hmm and multi-class adaboost," in European conference on computer vision, 2006: Springer, pp. 359-372.
[41] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in European Conference on Computer Vision, 2016: Springer, pp. 816-833.
[42] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data," in AAAI, 2017, pp. 4263-4270.
[43] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in CVPR, 2017.
[44] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," arXiv, no. Mar, 2017.
[45] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," Pattern Recognition, vol. 68, pp. 346-362, 2017.
[46] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4D human-object interactions for joint event segmentation, recognition, and object localization," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, pp. 1165-1179, 2017.
[47] S. Althloothi, M. H. Mahoor, X. Zhang, and R. M. Voyles, "Human activity recognition using multi-features and multiple kernel learning," Pattern recognition, vol. 47, no. 5, pp. 1800-1812, 2014.
[48] J. Liu, A. Shahroudy, D. Xu, A. K. Chichung, and G. Wang, "Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
[49] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," arXiv preprint arXiv:1801.07455, 2018.
[50] J. Dean. "Building Intelligent Systems withLarge Scale Deep Learning." Available: https://zh.scribd.com/document/355752799/Jeff-Dean-s-Lecture-for-YC-AI.
[51] S. Ramasamy Ramamurthy and N. Roy, "Recent trends in machine learning for human activity recognition—A survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1254, 2018.
[52] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," Journal of Big data, vol. 3, no. 1, p. 9, 2016.
[53] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," Remote sensing of Environment, vol. 62, no. 1, pp. 77-89, 1997.
[54] G. James, D. Witten, T. Hastie, and R. Tibshirani, An introduction to statistical learning. Springer, 2013.