PRACTICAL ALGORITHMS FOR VISION-BASED HUMAN ACTIVITY RECOGNITION AND HUMAN ACTION EVALUATION

YU XINBO

PhD

The Hong Kong Polytechnic University

2020

The Hong Kong Polytechnic University

Department of Computing

Practical Algorithms for Vision-Based Human Activity Recognition and Human Action Evaluation

YU XINBO

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

May 2020

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

____________________ (Signed)

____YU XINBO____(Name of Student)

Abstract

Human Activity Recognition (HAR) and Human Action Evaluation (HAE) are two main tasks of human activity analysis addressed in this thesis, which could be applied to many application domains like healthcare and physical rehabilitation, interactive entertainment, and video surveillance. These applications could alleviate the increasingly serious problem of population aging by bring improvements to the people's quality of life.

Existing HAR methods use various sensors like vision, wearable, and ambient sensors. With a comprehensive understanding of these sensors, this thesis focuses on the vision-based HAR. To test the effectiveness of existing methods, we collect a small real-world Activities of Daily Living (ADLs) dataset and implement some representative skeleton-based methods. We also propose an HAR framework called HARELCARE for developing practical HAR algorithms. Within the proposed HARELCARE framework, two effective HAR algorithms are developed and tested on the collected ADLs dataset. One of them is based on feature extraction, while the other is based on transfer learning. The results show both methods significantly outperform existing methods on our real-world ADLs dataset.

Not only tackling with small datasets, we also propose a Model-based Multimodal Network (MMNet) to handle HAR with increasingly larger public datasets. Since most of public datasets are collected with Kinect sensors, multiple data modalities like skeleton and RGB video are available. However, it remains a lack of effective multimodal methods that could further improve the existing methods. Our MMNet fuses different data modalities at the feature level. With extensive experiments, the proposed MMNet is proved effective and achieves the true state-of-the-art performances on three public datasets NTU-RGB+D, PKU-MMD and Northwestern-UCLA Multiview. The results of our HAR algorithms show great potential of our methods to be applied to wide applications.

Unlike HAR that focuses on activity classification, HAE is concerned with making judgements about the abnormality and even the quality of human actions. If performed effectively, HAE based on skeleton data can be used to monitor the outcomes of behavioural therapies for Alzheimer disease (AD). To do so, we propose a two-task Graph Convolutional Network (2T-GCN) to represent the skeleton data for both HAE tasks of abnormality detection and quality evaluation. It is first evaluated on the UI-PRMD dataset and found to perform well for abnormality detection. While for quality evaluation, in addition to the laboratory-collected UI-PRMD, we test it on a set of real exercise data collected from AD patients. Experimental results show that the numerical scores for some exercises performed by AD patients are consistent with their AD severity level assigned by a clinical staff. This shows the potential of our approach for monitoring AD and other neurodegenerative diseases.

Publications arising from the thesis

1. Bruce X.B. Yu, Keith C.C. Chan, Discovering Knowledge by Behavioral Analytics for Elderly Care", 2017 IEEE International Conference on Big Knowledge (ICBK), 284-289
2. Bruce X.B. Yu, Yan Liu, Keith C.C. Chan, Vision Based Daily Routine Recognition for Healthcare with Transfer Learning, 2020 (ICDHC)
3. Bruce X.B. Yu, Yan Liu, Keith C.C. Chan, Skeleton-Based Detection of Abnormalities in Human Actions Using Graph Convolutional Networks, IEEE TransAI 2020
4. Bruce X.B. Yu, Yan Liu, Keith C.C. Chan, A Survey of Sensor Modalities for Human Activity Recognition, 12th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2020)
5. Bruce X.B. Yu, Yan Liu, Keith C.C. Chan, Effective Human Activity Recognition Based on Small Datasets, 2020 International Association for Pattern Recognition (ICPR) (Submitted)
6. Bruce X.B. Yu, Yan Liu, Keith C.C. Chan, Two-Stream Graph Convolutional Network for Skeleton-Based Human Action Evaluation, 35th AAAI Conference on Artificial Intelligence (Submitted)
7. Bruce X.B. Yu, Yan Liu, Keith C.C. Chan, Teacher-Student Network: A Model-Based Multimodal Fusion Method for Action Recognition, 35th AAAI Conference on Artificial Intelligence (AAAI) (Submitted)
8. Bruce X. B. Yu, Yan Liu, Keith C. C. Chan, Qintai Yang, Xiaoying Wang, Skeleton-based human action evaluation using graph convolutional network for monitoring Alzheimer's progression, Pattern Recognition Journal (Submitted)
9. Bruce X.B. Yu, Yan Liu, Keith C.C. Chan, A Survey for Multi-modal Fusion Approaches for Activity Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (in preparation)

Acknowledgements

First, I must thank Prof. Keith C.C. Chan, for his endless guidance, inspiration, and encouragement throughout my life as a postgraduate student. I am greatly indebted to Prof. Chan for all that I learned from him both in terms of critical thinking and research style. Without his instant help and patient support, this thesis could never be possible to complete. I also very appreciate the numerous discussions and collaborations with the members in Prof. Chan’s research group.

Another person who had a large impact on my work is Dr. Yan Liu. Dr. Liu always cares about my research and daily life during the last year of my study. Although it is just one year, I cannot thank enough for her patient guidance. I am also grateful for the members in Dr. Liu’s research group. Working with them was enjoyable and enriching experience that brings me huge positive impact and interest in research.

I also would like to thank Prof. You Jia Jane, Dr. William K. W. Cheung, and Prof. Du Zhang for their valuable and insightful comments to this thesis, which inspires and motivates me a lot for the future direction of research.

My life as a postgraduate student would have been poorer without the hall life and fellow students in the PolyU. For their friendship and for many memorable experiences in many events and activities, I would like to thank Dr. Neil, Prof. Cecilia, Dr. Kee, Ruifeng and other friends in PolyU.

The data collection in Chapter 3 was under the help of my family members and the nursing home manager Miss Zhu. I could never pay back their priceless and strong support. I would dedicate this thesis to my family, and brothers and sisters.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Human Activity Recognition (HAR) and Human Action Evaluation (HAE) are two core tasks of human activity analysis [1]. HAR is concerned with recognizing different human activities from sensor data, while the goal of HAE is to evaluate the quality of a single activity performed by human subjects. It is important that these problems are well tackled because of the many applications they can have including physical rehabilitation, healthcare, video surveillance, and interactive entertainment. With the popularity of vision sensors, this thesis is tackling with practical algorithms for HAR and HAE from skeleton data and video sequences.

1.1 Background and Motivation

According to the latest world population prospection, the proportion of older people is predicted to reach nearly 1 billion in 2030, 1.5 billion in 2050 and it could reach nearly 2.9 billion by 2100. The population aged 80 or over is growing even faster than that of aged over 65. There were just 54 million people aged 80 or over worldwide in 1990, which is nearly tripled to 143 million in the year of 2019. Globally, the number of people aged 80 or above will be nearly tripled again to 426 million in 2050 and to further increased to 881 million in 2100 [2]. Aged people will have higher risks of suffering from various diseases which will even lead to death. These diseases, categorized as Noncommunicable Diseases (NCDs), like heart disease, diabetes, cancer, chronic lung disease, stroke, are together responsible for almost 70% of all deaths in the world [3]. With such a background, numerous healthcare applications emerged recently, which follows a similar process of HAR and HAE which includes steps like sensor selection, algorithm design, and recognition. The development of such HAR and HAE based applications could be formalized as a machine learning process as illustrated in Figure 1.1.

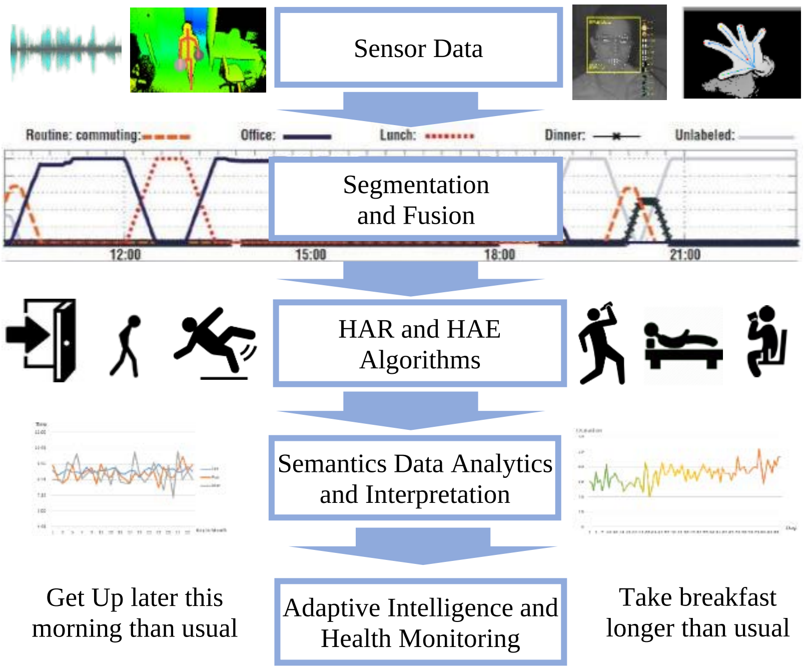

Sensor Data
Routine: commuting:
Office:
Lunch:
Dinner:
Unlabeled:
Segmentation and Fusion
12:00
15:00
18:00
21:00
HAR and HAE Algorithms
Semantics Data Analytics and Interpretation
Get Up later this morning than usual
Adaptive Intelligence and Health Monitoring
Take breakfast longer than usual

Figure 1.1 Procedure for HAR and HAE based healthcare applications

The HAR application categories include surveillance environments, entertainment environments, and healthcare systems, among which the healthcare system has two types namely daily life activity monitoring and rehabilitation applications [4]. Surveillance systems mostly aims to automatically detect and track human subjects in such a way to support security guards to monitor activities, which results in detecting suspicious activities and recognition of criminals. Detailed introduction of the surveillance application domain could be found in [4]. Literatures focusing on the recognition of entertainment activities such as sports [5], dance [6] and gaming [7] are usually related with the domain of healthcare and rehabilitation. From our observation, healthcare [8] occupies the majority of HAR applications among other application domains like manufacture monitoring [9], and security surveillance [10]. Hence, in this thesis, we delve deeper into the HAR and HAE applications in the healthcare domain.

Although sensor based human activity analysis has received much attention recently, whether existing human activity analysis methods could be applied to real-world environments has rarely been investigated as most of existing jobs are tested with data collected with controlled setup and informed human subjects in the laboratory environment. Due to the fact that most algorithms have been evaluated mainly with datasets collected in laboratories, this work is motivated by the lack of algorithms that have been shown to be effective practically in any particular application domains. Those healthcare applications could be solved well if human activity could be intelligently analyzed by machines. Human activity analysis could be categorized to subtasks like HAR and HAE. HAR systems could be performed as a lifelog or Activities of Daily Living (ADLs) recognition tool in physical environments, by which machines could have better ability of human behavior understanding. For ADLs recognition, there have been some effort to investigate into the recognition scenarios such as cooking [11], bathing [12], medication intake [13]. Whereas HAE could be essential for inferring abnormality from the actions or exercises performed by the subjects. Research in the past has focused on developing HAE solutions to determine abnormality in some specific activities such as walking imbalance [14], falling [15], sitting down and standing up [16]. 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. For example, when activities like falling or walking problems are detected or other NCDs in their late stages, such solutions might be useful for diagnosis through detecting activities but they could not be sufficient for elderly patients' lifestyle management which is essential for maintaining their independence.

Motivated by alleviating the urgent issues caused by the world wide population aging, we developed practical HAR and HAE algorithms and validated them on representative benchmarking datasets and also our own datasets collected from two real-world environments: home and nursing home. For the home environment, we use vision-based HAR to monitor the ADLs of an independent elderly, which could be used to infer the independence level of the elderly. The second one is using skeleton-based HAE to evaluate the morning exercise performance of the elderly people in a nursing home that has Alzheimer subjects, which is of great potential to provide evidence support for behavioral diagnosis and physical therapies.

1.2 Problem Definition

The above introduced applications can be tackled easily if problems from human activity analysis tasks are solved. In the following of this section, we introduce problems from the perspectives of the goals of HAR and HAE, their data characteristics and the related algorithms.

HAR is concerned with recognizing movements or actions of a human subject. Movements are often typical activities like drinking, sitting, standing, walking, and running that are conducted in indoor environments. They may also be more focused activities such as activities happened in a car or in a kitchen. The sensor data may be recorded remotely, such as radar, video, or other wireless sensors. Alternatively, data may be recorded by using wearable devices that have accelerometers and gyroscopes [1]. Currently, vision-based HAR problems could be focusing on different factors like, indoor or outdoor, fine-grained activity or general activities, single view or multiple view, and single subject or multiple subjects. With a definition of activity complexity, and investigation of the requirements in the healthcare domain, we focus on the indoor fine-grained HAR.

HAR algorithms could be focusing on single sensor modality like video [17], accelerometers and gyroscopes [18], and RFID [19]. Alternatively, it could also be tackled with multiple sensor modalities like video and accelerometers [20], and even various vision modalities from depth sensor [21]. Given a single sensor modality like an RGB video of particular duration that captures a series of activities, the raw video can be considered as a set of two-dimensional tables that contains millions of pixels with Green, Blue and Red attributes. It is commonly accepted that tasks like human detection, object detection could be performed on a single RGB frame. However, to infer the human activity from a video is far more complex than such simple tasks, which is still not well tackled recently although some large RGB video activity datasets are released. The RGB video might also lack of depth information from its single frame, which makes it insufficient to effective learning algorithms. Hence, multimodal solutions that could take the complementary advantages of different data modalities are a booming area in recent years [22]. However, how to fuse differently data modalities has seldom been well tackled. Existing multimodal methods usually fuse the results of different data modalities at the final layer of a DL model or at the fully connected layer, which could have subtle performance improvement and even decrease the overall HAR performance in some occasions, which is apparently not fully utilizing the complementary knowledge from different data modalities. In this thesis, we utilize both single modal and multimodal algorithms to improve the performance of vision-based HAR. In such a way, our algorithms could be practical for realistic application scenarios.

Another human activity analysis task is HAE that is closely related to the field of motion quality assessment, which aims to design evaluation methods to automatically assess the quality of a specific human motion. HAE relies on human tracking, human motion recognition, action segmentation, and effective methods for assessing the action quality [1]. As for HAE, most existing methods are tackled with logistic regression algorithms that classify activities into the binary classes as normal and abnormal. These algorithms do not give continuous numerical evaluation scores and as a result, activities that are performed while a patient is in rehabilitation may not be easily determined. Also, the algorithms that have been used are trained based on the use of laboratory data collected on young subjects [23] [24]. The performance of these algorithms in more realistic environment may not be very accurate. In addition to this problem, existing HAE methods are developed based on motion sensors like Kinect or other motion capture sensors. The issue of which exercises are good for inferring abnormality has not been well tackled.

1.3 Overview of Solutions

There are plenty of sensor-based behavior analysis technologies and applications emerged in recent years like vision sensors, ambient sensors [19] [25] [26] and wearable sensors [18] [27]. With a thorough investigation of the advantages and disadvantages of different sensors in Chapter 3, we found depth motion sensors that remarkably attract the interest of researchers could be capable to tackle with human activity analysis tasks. Precisely, we adopted a depth sensor known as Kinect v2 to handle both HAR and HAE problems. As shown in Figure 1.2, Kinect v2 has three types of sensors namely depth sensor, RGB camera, and audio sensors. Two datasets were then collected by using Kinect v2 in two realistic environments, which is introduced in Section 3.4.2 and Section 3.4.3.

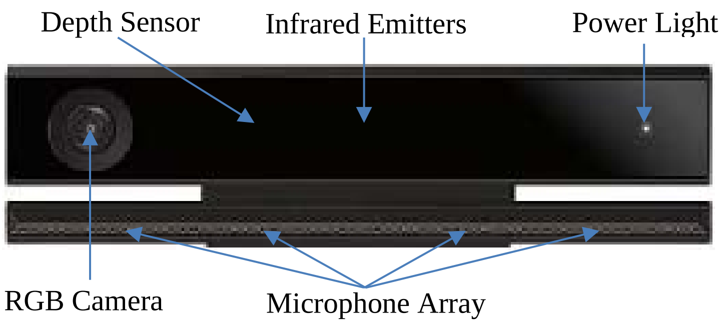


Figure 1.2 Kinect v2 for windows sensor

To design practical HAR algorithms for logging ADLs, we proposed an HAR framework that could accommodate different HAR algorithms and that could work with one or more data modalities. Two algorithms for real-world scenarios when large training data are not available are proposed from the HAR framework for learning patterns from the data collected in such a way that the first algorithm focuses on the skeleton modality based on feature extraction. The second algorithm also focuses on the skeleton modality but based on transfer learning.

To better tackle the HAR problem with data driven methods, we proposed an algorithm that can handle multimodality by fusing the skeleton and video modalities at the feature level. We conduct experiments on some large public datasets to verify the effectiveness of our multimodal method and found it achieved state-of-the-art performances, which indicates the great potential of our solutions to be utilized in logging ADLs of the independent elderly to provide supportive information for healthcare and remote care services.

The HAE algorithm is based on the skeleton modality for the evaluation of physical exercises of patients. To test our HAE algorithm, we conducted several experiments on a benchmarking dataset and achieved good performance. To determine which activities are good for the prediction of abnormality, we collected a morning physical exercise dataset from a nursing home where there are a number of elderly subjects who are suffered from Alzheimer's disease (AD). We conducted abnormality analysis for the exercises on the dataset with our algorithm with a continuous numerical score assigned to different activities based on how much they deviate from abnormality. The experimental results show that our proposed algorithms can have great potential for practical applications.

1.4 Contributions

The contribution of this thesis comes from practical and algorithmic perspectives. For the practical perspective, we first defined different levels of the human activity complexity as a standard for evaluating the capability of different sensors and existing HAR tasks. Second, we reinvestigated various sensors and made a comprehensive understanding of their specific arrangement and feasibility for human activity analysis. Third, we filled the gap between the vision-based activity analysis and healthcare domain by coming up with two real-world application scenarios, which could be further investigated based on our experimental results. Precisely, we collected two datasets in two different real-world scenarios. One is an ADLs dataset collected from the bedroom of an independent elderly person. The second one is a morning exercise dataset collected from a nursing home that has Alzheimer subjects. We then designed a framework named HARELCARE that accommodates various HAR and HAE solutions with different components to choose from in each stage of the framework.

For algorithmic contributions, we proposed four algorithms for human activity analysis. Three different HAR algorithms that covers both single modal and multimodal HAR methods are proposed for both effectiveness and practical purposes. For single modal algorithms, we proposed a traditional feature extraction method that is easy to implement and a transfer learning-based method for scenarios when large training data is not available. For the multimodal algorithm called Model-based Multimodal Network (MMNet), we utilized the skeleton modality and RGB modality and proposed to fuse them at feature level. With extensive experiments, the proposed MMNet consistently achieves state-of-the-art accuracies on three public datasets NTU-RGB+D [21], PKU-MMD [28] and Northwestern-UCLA Multiview [29].

Besides HAR algorithms, we also proposed an HAE algorithm called two-task Graph Convolutional Network (2T-GCN). The results not only outperform existing methods with their criteria on a benchmarking dataset called UI-PRMD [24], but also indicate that Kinect v2 is more capable than the Vicon motion capture for HAE. It is then applied to our nursing home morning exercise dataset. The results show encouraging abnormality prediction performance and high consistency with clinical evaluation of AD, which indicates the great potential of our HAE algorithm for supporting clinical criteria of Alzheimer diagnosis and behavioral therapies with objective evidence.

1.5 Organization of Thesis

In Chapter 2, a comprehensive review of the related literature that sheds light on single modal and multimodal HAR algorithms is introduced. We also present a description of existing work on HAE.

In Chapter 3, we describe the sensor selection and arrangement of sensing environments for human activity analysis based on a proposed standard definition of activity complexity and an investigation of the capabilities of different sensors. Besides, we also introduce the data collection methods and the two collected datasets: ADLs dataset and nursing home morning exercise dataset in this chapter.

In Chapter 4, we introduce the proposed HARELCARE framework and two HAR algorithms derived from the framework. The experimental results of the two HAR algorithms on our ADLs dataset and discussion are presented in the same chapter.

In Chapter 5, the MMNet model is introduced and tested on benchmark datasets like NTU-RGB+D, PKU-MMD and Northwestern-UCLA Multiview. We also compared experimental results of this multimodal algorithm with the other two algorithms introduced in Chapter 4 on our ADLs dataset.

In Chapter 6, we introduce the HAE algorithm called 2T-GCN. Experimental results of our HAE algorithm on the benchmarking dataset called UI-PRMD and our nursing home morning exercise dataset are also presented.

Finally, in Chapter 7, we summarize the thesis and discuss whether the proposed algorithms could be used for human activity analysis. The limitations of our proposed algorithms and directions of future job are also concluded.

Chapter 2

Related Work

During recent years, vision-based human activity analysis has a decent number of methods emerged. In this chapter, we survey the related literature of existing algorithms for vision-based human activity analysis from single modal to multimodal methods, and from HAR to HAE tasks with our critical and practical concerns.

2.1 Single Modal HAR

Single modal vision-based methods for HAR have two rough groups: representation based on local features [30], [31] and the body skeleton [32], [21]. HAR systems that base on local features are independent to the choice of sensors as they only use raw depth data and more robust to occlusion since depth camera is usually installed on the ceiling. Comparing with approaches based local features, with a skeleton or pose extraction layer on raw depth data or RGB data, skeleton representation of human body significantly alleviates the complexity and computation cost of human activity analysis tasks. Most of existing algorithms usually focus on a specific data modality of the datasets. In the remaining of this section, we first introduce current algorithms proposed to tackle the skeleton modality and then describe the algorithms for the RGB video modality.

2.1.1 Skeleton-Based HAR

The skeleton feature could be available from off-the-shelf vision sensors like depth cameras, Motion Capture (Mocap) systems, and RGB cameras. Depth motion sensors like Kinect and RealSense and even RGB cameras have the ability to detect human body skeletons. Since the release of such motion sensors, skeleton-based HAR is a booming area in computer vision. A bunch of large datasets like MSRDailyActivity3D [32], Multiview RGB-D Event Dataset [33], UWA3D Multiview II [34], NTU RGB+D [21], PKU-MMD [28], UESTC RGB-D Varying-view action [35], and NTU RGB+D 120 [36] were collected by using Kinect v1 and v2 sensors. Among these datasets, the number of activities and subjects involved in the NTU RGB+D [21] dataset is the largest. Activities in NTU RGB+D are grouped into three categories: daily actions, mutual actions, and medical actions. Meanwhile, UESTC RGB-D Varying-view action dataset was designed for human-robot interaction includes 40 sports related actions and involves 118 subjects.

As skeleton data of an activity basically has the sequential character, traditional algorithms like DTW [37], HMM [38], and SVM [39] are commonly used, which is then dominated by DL algorithms [40]. Wang et al. [40] reviewed DL models for HAR tasks, which includes Convolutional Neural Network (CNN), Deep neural network (DNN), , Stacked autoencoder (SAE), etc. Existing researches of skeleton-based HAR mainly focus on three directions for the improvement of activity recognition. The first direction focuses on data preprocessing and data cleaning. For example, Liu et al. [41] introduced an algorithm to remove the noise of skeleton joint by learning a model that reconstructs more accurate skeleton data. Similar jobs have been proposed by Pengfei et al. [42]. The second approach improves the HAR benchmarks by proposing novel learning or representing models. Liu et al. [43] introduced a context aware LSTM model that could learn which part of joints contribute to the HAR. Since the induction of ST-GCN [44] some enhanced versions of GCN models have been proposed that improve the ST-GCN by considering other physical prior knowledge. For example, Li et al. [45] tries to model discriminative features from actional and structural links of the skeleton graph. Except GCN, motivated by cooccurrence learning, Chao et al. [46] proposed the hierarchical cooccurrence network (HCN) that learns point-level features aggregated to cooccurrence features with a hierarchical methodology. The co-occurrence features refer to the interactions and combinations of some subsets of skeleton joints that characterizes an action [47]. Considering both the graph and cooccurrence characteristics, Si et al. [48] proposed AGC-LSTM that achieved high accuracy on the NTU-RGB-D dataset. Similarly, the (directed graph network (DGN) [49] this is a good introduction of an acronym) achieved the better performance than the AGC-LSTM on the NTU RGB+D dataset with a smaller margin. Existing DL models could successfully capture the spatial and temporal features well on the skeleton modality, but to further improve the performance, appearance features from the RGB modality might be useful for preventing existing skeleton-based DL models from overfitting. Hence, we will further investigate the multimodal algorithms in Section 2.3.

2.1.2 RGB Video Based HAR

Traditionally, methods that utilized local features based approaches are proposed for recognizing simple activities like fall and hand gestures [30]. Another job proposed by Elangovan et al. [31] attempted to use a local feature based algorithm to analyze three types of interactions between human to object, human to human, and human to vehicle, which is at a very rough and casual HAR level and has little generalization ability as its features are fixed. Although traditional methods could not provide applicable fine-grained HAR solutions, a vision-based approach proposed by Albert et al. [50] for monitoring hand hygiene compliance in hospital outperforms the accuracy of three people covert observation, which identifies the potential of vision-based sensor for human activity analysis tasks.

Other than local features based approaches, data driven approaches have also been proposed with large RGB video datasets like DeepMind Kinetics-400 and Kinetics-600 collected from YouTube videos by Google [51] that contain much more activity classes with the number of 400 and 600 respectively. Kinetics datasets insist the philosophy of accommodating as many activity categories as possible, by which could be applied to video search, surveillance and robotics. Other similar large scale complex video activity datasets like UCF-10 [17], HMDB-51[52], Sports1M[53], ActivityNet [54], Multi-THUMOS [55] have revitalized the domain of HAR and inspired new ideas and research directions.

Carreira proposed I3D [56] that uses pre-trained inflated Inception-V1 mode miniKinetics as back bone to improve the performance of UCF-101 and HMDB-51 by end-to-end fine-tuning. Two data modalities RGB and optical flow, extracted by the TV-L1 algorithm are used, as in their two-stream model. It turns out that the optical flow modality performs better for UCF-101 and HMDB-51 dataset but is surpassed by RGB modality on miniKinetics. On top of I3D, Xie et al. [57] considers about the speed-accuracy trade-offs for video classification, and proposed a S3D model that further improves the performance of [58]. It is worth mentioning that the S3D [57] was implemented on 56 GPUs with thes batch size set to 6 for each GPU, which reflects the huge computational cost of video-based HAR. It is intuitive that fusing the results of S3D with skeleton-based methods could boost the recognition accuracy. However, from our observation, the S3D will stick at around 12% for the Top-1 accuracy for NTU RGB+D. This might be due to the fact that S3D is designed for datasets like UCF-101 and Kinetics that are different with the NTU RGB+D. Precisely, activities in UCF-101 and Kinetics cover broader activity categories of both indoor and outdoor. The indoor activities in NTU RGB+D might be more challenging for such video-based methods. Another reason might be due to the limitation of computational resources as a workstation with 56 GPUs is not common.

In terms of application domains like healthcare, the level or resolution of those activities in the Kinetics dataset might not enough for fitting their application requirements. By focusing on a particular scenario, Sigurdsson et al. [59] introduced a crowdsourcing approach named Hollywood in Homes to collect a Charades dataset that is annotated with free-text descriptions and labeled with 157 action classes 46 object classes. Sigurdsson et al. [60] further examine whether kinds of information like activity labeling, temporal extent will be useful to achieve substantial gains in HAR from the perception perspective of human brain. Finally, they suggest that fine-grained classification of activities that share similar label of objects and verbs is essential for accuracy improvement and better activity understanding, which indicates that RGB data alone is faced with great challenge for the indoor human behavior understanding.

2.2 Multimodal HAR

It has been generally accepted that multimodal HAR approaches have the potential to improve recognition and could be capable of distinguishing difficult activities. The multimodal fusion analysis of [61] for the Opportunity dataset indicates that feeding more data channels to its proposed DeepConvLSTM would deliver improved performance. Similarly. the experimental results of [36] for NTU RGB+D 120 dataset also indicate that extra data modalities contribute to classification accuracy. Multimodal HAR could be roughly categorized into two classes: vision-based multimodal [62], [63], [64] and vision-wearable-based multimodal [65], [66], [61]. Algorithms for multimodal HAR share the similar trend that uses DL to extract discriminative features. Baltrušaitis et al. [22] summarized five technical challenges of multimodal solutions: representation, translation, alignment, fusion, and co-learning. The key issue of multimodal methods is to find proper ways of data fusion with the co-learning concept in mind.

The 4DHOI model proposed by Wei et al. [62] attempts to represent both contextual objects and 3D human poses in events by using a spatial temporal graph with hierarchical structure. The fusion concept of [64] has two approaches namely intermediate fusion and late fusion. For the intermediate fusion approach, the skeleton and RGB modalities are separately pretrained first, then a shared representation was generated with a concatenation of their high-level representations. Whereas the late fusion scheme simply combines the results of two modalities. Pan et al. [67] proposed a cross-stream selective network (CSN) that leverages the correlation and complementarity of different input streams. CSN is designed to find the most discriminative temporal frames aligned to spatial frames and globally endows different weights to RGB and optical flow groups. Unlike CSN, our method will select which body part from the RGB stream will provide extra discriminative information. This concept has been attempted by [68], which achieved decent improvement by using an attention mechanism that focuses on two hands. However, [68] might neglect some activities with human-object interaction that involves lower body like “put on shoes” and take off shoes. Other similar jobs also have been proposed in [69] and [70]. In our method, we utilize the effective attention features from the skeleton modality that focus on more body areas including the head and feet instead of just two hands as in most existing jobs in [68], [69] and [70].

2.3 Skeleton-based HAE

Very few jobs have reviewed the benchmarks of HAE. Ahad et al. [71] briefly collected some HAE datasets for healthcare, but most of the included datasets are focusing on HAR instead of HAE. Meanwhile, as far as we know, there is no job that investigates the standard evaluation method of HAE algorithms. Brook et al. [72] introduced the performance of the Kinect v1 for measuring movements of people who have Parkinson’s disease. It turns out that the Kinect v1 sensor scan accurately capture the gross timing and spatial features of body movements that is relevant to some clinical criteria. Kinect has also been used for cognitive stimulation for dementia individuals in [73]. The KiMentia system proposed by [73] provides an user interface with some simple HAR functions and its concept is commonly supported by clinical experts. Many researches have been conducted by using the Kinect for measure the wellness or accuracy of actions. Hence, we further investigate the evaluation methods of some latest benchmarks that use Kinect sensor to collect skeleton data as listed in Table 2.1.

Ortega et al. [23] proposed to use the Kinect v1 to monitor psychomotor exercises like “touch the left eye with the right hand”, “Touch the right eye with the right hand.”, and “raise the right hand” etc. Although the system of [23] could detect 14 psychomotor exercises with an accuracy of 96.28%, it did not perform quality evaluation for the exercises. Tao et al. [74] proposed a model for online motion quality evaluation and validated on the SPHERE dataset which includes three sub- datasets, Staircase2014, Walking2015 and SitStand2015. The SPHERE dataset is originally collected for a competition and just provide the body center data instead of the whole skeleton. Tao et al. [74] compared various Hidden Markov Model (HMM) models which is traditional, and its performance has been beaten in various datasets by DL models. Vakanski et al. [24] collected a fitness exercise dataset named UI-PRMD for HAE algorithm evaluation. Liao et al. [75] proposed a DL framework to encode the skeleton data of the UI-PRMD dataset, which is supervised by a quality score function. Since the UI-PRMD dataset did not provide a standard evaluation method, Liao et al. [75] proposed its own quality score function, which makes it meaningless to train a representation model to fit it as the results could already be inferred by the quality score function. Lack of standard evaluation method also makes the dataset hard to be compared by a similar job in [76]. Unlike the 10 incorrect exercises in UI-PRMD that are simulated by the ones that perform the other correct motion sequences, Antunes et al. [77] collected a dataset AHA-3D that is performed by both elderly and young people but it is not targeting to any specific disease. Meanwhile, by the date of this job, the AHA-3D dataset is not publicly accessible. The evaluation method in [77] is per frame but not in terms of the whole action sequence. Elkholy et al. [16] collected a dataset that is similar with SPHERE [74] and proposed a similar HMM based method that has less computational overhead than the one proposed by [74]. The training process of [16] is supervised by the score of abnormality degree (on the scale of 1 to 5) evaluated by professional specialist.

Table 2.1 Benchmark Human Activity Recognition Datasets (NS Stands for Number of Subjects, NA Stands for Number of Activities, and NR Stands for Number of repetitions)

#	Dataset	Year	Sensor	Disease	NS	NA	NR
1	Ortega et al. [23]	2014	Kinect v1	cognitive damage and deterioration	15	14	Psychomotor exercises
2	SPHERE [74]	2016	Kinect v2	Stroke, PD	10	3	48/40/109
3	UI-PRMD [24]	2018	Vicon, Kinect v2	Rehabilitation	20	10	1326
4	AHA-3D [77]	2018	Kinect v2	Not mentioned	23	4	NULL
5	EJMQA [16]	2020	Kinect v2	neuromusculoskeletal disorders	32	4	NULL
6	Our Nursing Home Dataset	2020	Kinect v2	Alzheimer's disease	25	6	869

To measure the abnormality and action quality, exercise or action representation models are usually trained on normal action sequences and then tested on new observations to infer anomalies according to the similarity of the new observation generated by the model. To the beset of our knowledge, there is no application that utilizes HAE to support real disease diagnosis and treatment. Some trial projects have been developed that base on gaming scenarios like bowling in the Kinect Project [78] and cognitive stimulation as in the KiMentia system [73]. According to the above analysis of the state-of-the-arts methods, there are some challenges and issues need to be tackled for the popularization of HAE based healthcare applications. First, the datasets need to be naturally collected instead of simulated by young subjects and performed in laboratory environment. Second, there is no standard evaluation methods being developed for action assessment although some benchmark datasets have been collected. Third, once effective evaluation methods have been developed, it needs to be validated in real treatments that follows clinical validation procedures.

Chapter 3

Vision Environments for HAR and HAE

In this chapter, we elaborate the activity complexity definition standard and a comparison of the capability of various sensors for HAR and their corresponding arrangements in the first three sections. In Section 3.4, we introduce the two datasets collected in real-world environments.

3.1 Definition of Activity Complexity

From the activity perspective, a clear definition of the human activity complexity is crucial for evaluating the HAR capability of different sensors. It is challenging to concretely group human activities as they are complex in terms of many structural characteristics like different levels of activities in a hierarchical structure, various activity durations, different locations, and the numbers of people and objects involved. Previous research defined activity complexity by only considering the time span as shown in Table 3.1 [79], which might not adequate to represent attributes of human activities. Low-level of activity recognition such as human subject tracking and body posture analysis was covered by Aggarwal and Cai [80]. Some performance-oriented jobs usually verify their models' accuracy on benchmark datasets with the duration of the recognized activities limited to seconds. For ease of explanation, we concentrate the case of single subject HAR since multi-subjects activity recognition, to some extent, could be expanded from single subject HAR with human identification. Considering three aspects: object, time, and location, we come up with a space (see Figure 3.1) to represent activity complexity for single subject human activities, which accommodates the activity categories in a hierarchical structure. This activity complexity definition is used to evaluate sensor capability in Section 3.3. As far as we know, this might be the first study that examines the complexity of human activity and the required level of HAR for applications.

Table 3.1 Activity Levels Concerning Sematic Meaning and Duration

Activity Level	**Semantic Complexity**	**Duration**	**Activity Examples**
I	Gesture or pose	Frames to seconds	Hand gesture, human appearance
II	Action	Seconds to minutes	Sit down, stand up, fall down,
III	Activity	Minutes to hours	Writing, watching TV, sleeping, cooking
IV	Behavior	Hours to Days	Daily routine, morning routine
V	Concurrent activities	Seconds to hours	Cooperation, dispersion, planning

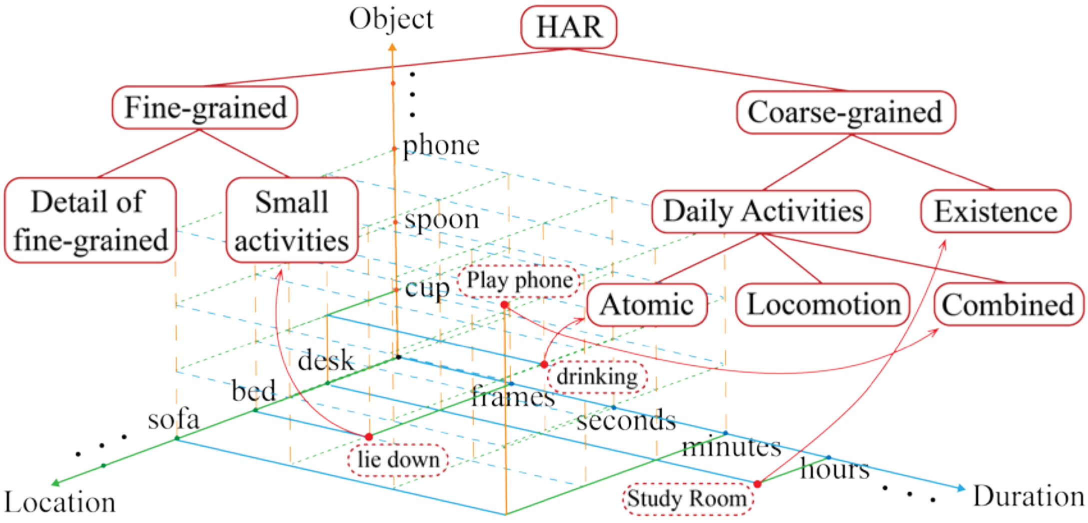


Figure 3.1 Our HAR categorization scheme

3.2 HAR Sensor Arrangements

Recent work in the development of HAR methods depends on the arrangements of different sensors. For ambient sensors, some RFID technologies, such as that reported in [19], require that RFID tags be 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 be embedded in the bed cloth. This approach is considered as non-intrusive [25]. The main disadvantage of RFID is that, due to the mutual interference of RFID signals when two objects are closely positioned, the detection accuracy will be affected by the high signal 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 activity recognition. In wireless communications, channel state information (CSI) is known as channel characteristics 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 [81]. It also has the advantage of cross wall sensing ability, but such an approach lacks a theoretical foundation that elaborates its capability for multi-users activity recognition. Wi-Fi CSI based HAR requires strict sensor positioning, which makes it hard to install and adapt to different real-world environments. Wi-Fi CSI remains at its early research stage for HAR and it has a lack of comparison with other sensors in terms of their measurement accuracy.

Wearable devices could be an appropriate choice for activity recognition. Since each sensor modality has its specific limitations, there has been some effort to fuse vision and inertial sensor data to try to improve HAR accuracy [65], [20]. In [82], a review of previous work that use both depth camera and inertial sensors to collect multimodal

3D data was presented, which provides a summary of the similarities in the features that are utilized for such sensor fusion approaches. However, inertial modality does not provide any context information for fine-grained (e.g. human-object interaction) HAR tasks. Besides, due to intrinsic battery limitation, this approach is considered too intrusive as batteries need to be replaced to allow wearable sensor devices to be capable for long-term activity monitoring.

When it comes to wearable devices, it is still unclear whether or not adding extra modalities can improve HAR accuracy. Based on an Opportunity dataset, the multimodal fusion analysis of [61] reveals that the more data channels are involved for its proposed DL model named DeepConvLSTM, the better HAR the recognition performance can be. For example, starting from a F_1 score of 69% that used only the accelerometers of the *Opportunity* dataset, the performance improved on average by 15% when fused accelerometers and gyroscopes and by 20% when fused accelerometers, gyroscopes and magnetic channels.

However, the use of different combinations of sensor modalities in experiments on the dataset named Berkeley Multimodal Human Activity Database (MHAD), the improvement on the performance is very limited when adding more data modalities (from around 98% to 100%) [20]. It is also concluded that adding extra modality may even lower the HAR accuracy, which means the extra modality could not bring more discriminative features. Besides, the increased problem complexity and decreased practical usability make the multimodal HAR that uses various sensors unpopular among end users and other stakeholders.

Table 3.2 Benchmark Human Activity Recognition Datasets (NS Stands for Number of Subjects, NA Stands for Number of Activities)

#	Dataset	Sensor	Sensor Type	NS	NA	Activity
1	MSRDailyActivity3D [32]	Kinect v1	Vision	10	16	Actions
2	PKU-MMD [28]	Kinect v2	Vision	66	51	Actions
3	NTU RGB+D 120 [36]	Kinect v2	Vision	32	120	Actions
4	Kasteren Dataset [25]	State change	Ambient	1	8	Daily activity
5	Freiburg Dataset [83]	Audio	Ambient	--	22	Daily activity
6	Smart Carpet Dataset [19]	RFID	Ambient	13	2	Fall detection
7	WiAR Dataset [26]	Wi-Fi	Ambient	10	16	Gestures, activities
8	PAMAP2 [18]	3 3-DOF IMUs	Wearable	9	18	Daily activities
9	Opportunity dataset [65]	IMUs, 72 sensors of 10 modalities	Multimodal	12	21	Morning activities
10	Berkeley MHAD [20]	Mocap, Kinect v1, camera, acc, audio	Multimodal	12	11	Actions

To compare the capability of different sensors, we collected representative publicly available datasets that use various sensor arrangement as listed in Table 3.2. It is note-worthy that the NTU RGB+D 120 [36] could be the dataset that includes data involving relatively more complex activities. This dataset is collected based on vision sensors. Based on an analysis of the activities in Table 3.2, it is noticeable that vision sensors are relatively more capable for HAR comparing with other ambient and wearable sensors that have been adopted when concern the number of subjects involved in their datasets, and the number of activity classes that they try to recognize. However, with the ambition to simultaneously recognize activities with different levels of activity complexities like varied activity resolutions, high- and low-level activities, and human-object interactions, the NTU RGB-D 120 [36] that contains a larger number of different activities has been made available for testing. So far, this goal has not been achieved too well as performance with the dataset suffers from relatively low accuracy. The most accurate recognition rate achieved was at around 65% [36]. Also, if ADLs recognition for NCDs is 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 increased deployment complexity, large datasets might not be feasible when developing HAR methods for real-world healthcare applications.

3.3 Sensor Capability Comparison

Table 3.3 summarizes sensors that are available in each sensor modality for HAR with their general advantages and disadvantages. Popular sensors used by researchers has been introduced in the last section.

Table 3.3 Sensors for Human Activity Recognition and Behavior Understanding

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 Light condition Pervasiveness Computational cost Privacy	Location limited Installation complexity Maintenance	Intrusive/obtrusive Acceptance of subject Battery life

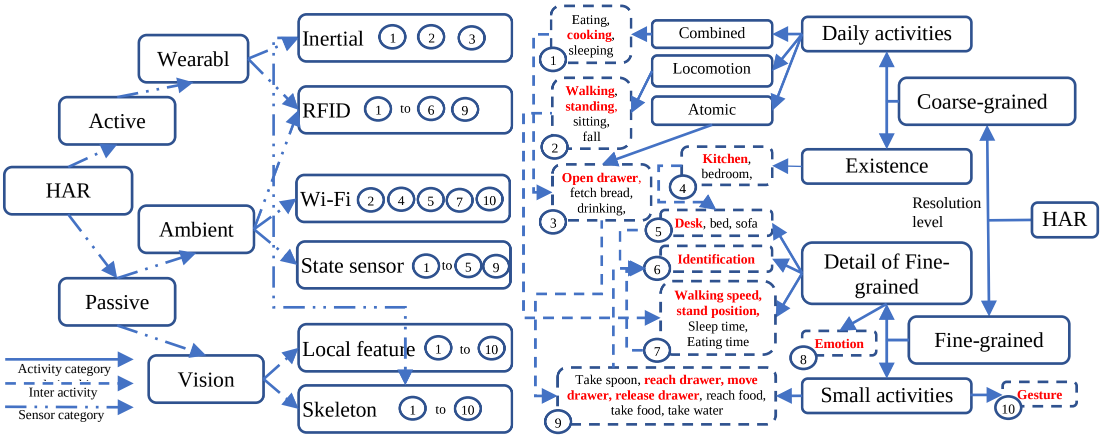


Figure 3.2 HAR modality hierarchical categories with relevant potential activity categories

Based on the reviewed representative benchmark datasets from each category of sensor-based HAR and the activity complexity definition as in Figure 3.1, the capabilities of different sensors are summarized in Figure 3.2. Based on such analysis, we could observe that the vision sensor is the most capable for HAR tasks (from activity categories 1 to 10). The activities highlighted in red in each activity category on the right side of Figure 3.2 are taking the example of breakfast preparation procedure from the job of [11].

One recent trend of the multimodal HAR is the fusion between inertial sensors and vision sensors as described in the multimodal HAR datasets in Section 3.2. However, according to various modality combination experiment results on Berkeley MHAD, the improvement of the performance by adding more data modalities is very limited (from around 98% to 100%) [20]. Sometimes, adding extra modality will even lower the HAR accuracy, which renders the extra modality in vain. The increased problem complexity and affected usability also make the multimodal HAR hard to be popularized among end users as well as other stakeholders.

Considering ambient sensors, take the RFID technology used by [19, 84, 85] for example, it needs to install RFID tag on the entire floor of a user’s living environment for the purpose of detecting if the user is near the bed or not with their claim that the cloth with RFID antenna is unobtrusive. RFID noise interference remains an issue if two objects are very closely located. Whereas, the use of state change sensors needs to install a quite number of sensors to all the related locations, but it could only do some coarse-grained activity recognition. Although Wi-Fi CSI emerged as a novel method that has the advantage of cross wall sensing ability, it remains lack of theoretical foundation and practical methodology for HAR.

Besides ambient sensors, wearable devices could be an appropriate choice for outdoor activity recognition. In a few existing jobs, it has been studied that fusing data of inertial and vision sensors could improve the HAR performance. A review by Chen et al. [82] summarized previous jobs that using both depth camera and inertial sensors to collect multimodal 3D data. Chen et al. gave the common features utilized with their fusion approaches. However, the inertial modality does not provide any context information for fine-grained HAR tasks like human-object interaction. Besides, due to the intrinsic battery limitation, it is intrusive for users as they need to wear sensor devices for long-term monitoring.

Comparing all the public available datasets, it is worth mentioning that NTU RGB+D 120 is by far the largest benchmark HAR dataset among all mentioned datasets concerning perspectives like subjects involved, number of activity classes, and number of viewpoints. Many algorithms have been proposed and tested on the on the NTU RGB+D dataset. Some of them model activities with spatial and temporal models by using CNN and LSTM algorithms [86], [87]. While some of them try to model the most informative joints for HAR with context-aware LSTM algorithm [43] or remove the noise of the skeleton data for view invariant recognition [88] [41]. Another potential method is using the contextual information to improve the HAR accuracy by modeling human-object interaction [62], which has no experimental result on the NTU RGB-D dataset.

3.4 Our Sensor Arrangement

3.4.1 Vision Sensor Selection

There are different vision sensors to choose from for tackling with the representation based human activity analysis. As Figure 3.3 shows, available vision devices that support 2D or 3D skeleton retrieval could have three types: Mocap system, depth camera, and RGB camera.

Mocap system companies like OptiTrack, Qualisys, and VICON provide such system for areas like biomechanics, sports, engineering, and entertainment. These Mocap systems can provide very accurate skeletal data but suffered from high price and low flexibility for commercialization purpose.

Depth cameras could be based on technologies like structured-light sensor or time-of-flight sensor. Some off-the-shelf commercial depth cameras like Kinect and Intel RealSense can retrieve skeleton data in a significantly affordable way with acceptable accuracy for HAR.

RGB cameras includes monocular camera and multiple cameras. The cheapest RGB cameras are also able to retrieve 2D skeleton [89] or 3D skeleton [90], [91], which require higher computational cost and might hinder the development of real-time prototype applications. Detecting 2D skeleton from RGB image resembles the COCO Keypoint Challenge [92].

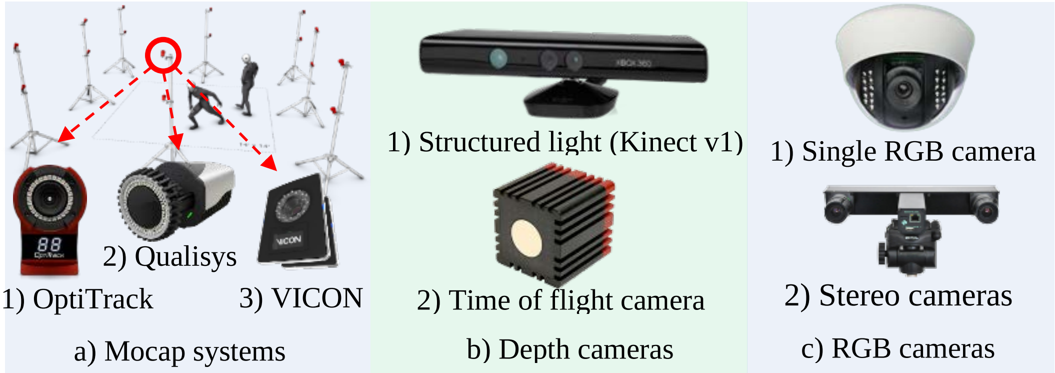


Figure 3.3 Vision devices that allow for 2D/3D skeleton retrieval

With the capability of collecting skeleton joints data, Kinect sensor dominated the vision-based HAR. According to the list of public benchmark datasets in the survey of Han et al. [93], there were 29 out of 41 datasets were collected by Kinect sensor. Mocap ranked as the second most popular approach in Han et al.'s survey. While RGB cameras is the least capable yet the most affordable vision sensor that could be used for the HAR.

Due to the popularity of depth cameras, we list all the off-the-shelf depth cameras in Figure 3.4 and make a comparison. The market of depth sensors could be classified into two groups: Microsoft and Intel group. In Figure 3.4, the Microsoft group has sensor a, b, c, and g, among which sensor g is an improved version of a, b, and c. Whereas the Intel group has released two sensor groups (based on range type in Table 3.4): d with advanced model h; e and f with advanced version j as shown in Figure 3.4.

Xtion Live Pro and Leap Motion (see Figure 3.4) have seldom been used in collecting datasets [94]. Kinect v1 and devices that are similar with Kinect v1 (which includes

Xtion Live Pro and Primesense Carmine) enables developers to retrieve skeleton joints data through Kinect SDK v1.8 [95] or OpenNI SDK v1 [96] libraries, respectively. With more advanced capability than sensor a, b, and c as shown in Figure 3.4, Kinect v2 is by far the most capable 3D devices for body skeleton joints retrieval through the Kinect SDK v2.0 [97]. Despite the popularity of Kinect sensors, Microsoft has discontinued the manufacturing of Kinect v2 in 2017 and brought the Kinect technology to its augmented reality project Hololens [98] which is similar with google Tango project. Fortunately, the shutdown rumor of OpenNI after acquired by Apple in 2014 was exaggerated, and its second version OpenNI SDK v2 was maintained by officially Occipital and available again for multiple platforms (including macOS, Windows, Linux, and Android) [99], which might make the HAR into a new era as improved accuracy of skeleton retrieval will intuitively improve the final HAR performance. The potential of Intel SR300 as shown in Figure 3.4 for hand joints and emotion retrieval have not been explored by researchers. With the active market of depth sensor, we could imagine that the future of HAR will rely on these technologies. All depth camera sensors and their corresponding SDKs are listed in Table 3.4, in which the usable range would be roughly divided into two categories as local (below 1.5 meters) and global (1.5 to 4.5 meters) as grouped on the right side of Figure 3.4. From the best of our experience, in terms of skeleton retrieval stability, wellness of support, and community activeness, Kinect v2 and RealSense SR300 might be the best choices for global and local data collection, respectively.

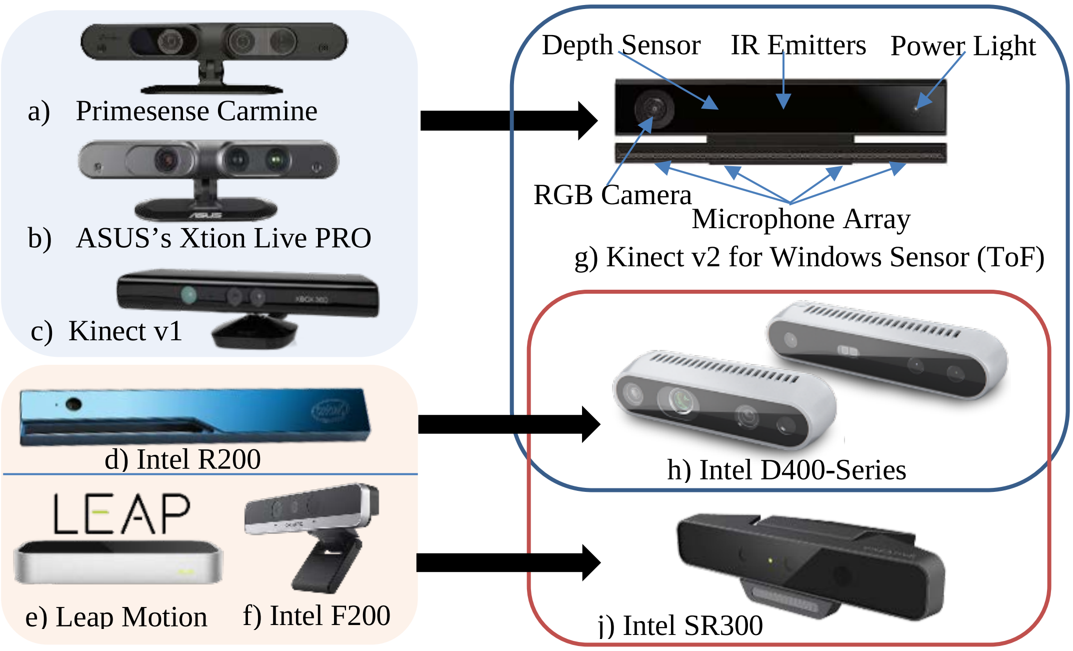


Figure 3.4 Off the shelf RGB-D cameras

Table 3.4 Comparison of Skeleton Retrieval SDKs

SDK	**Device**	**Year**	**Range**	**Body Part**
Kinect SDK v1 [95]	K4W v1	2010	0.8 - 3.5m	Body
Kinect SDK v2 [97]	K4W v2	2013	0.5 - 4.5m	Body, hand, face
OpenNI SDK v1 [96] and v2 [99]	K4W v1 Xtion Live Primesense	2010 2011 2013	0.8 - 3.5m 0.8 - 3.5m 0.35 - 3m	Body, hand
Leap Motion SDK v2 [100]	Leap Motion	2014	Up to 1m	Hand
Intel RealSense SDK v1 [101]	Intel F200 Intel SR300 Intel R200	2014 2015 2016	0.2 - 1.2m 0.2 - 1.5m up to 3-4 m	Hand, face Body, hand, face Up body, face
Intel RealSense SDK v2 [102]	Intel SR300, D400-Series	2016 2016	0.2 - 1.5m, 0.11 - 10m	Body, hand, face Up body, face

3.4.2 Home Environment

Amidst various NCDs, it is important that the elderly can maintain the ability to live independently and with dignity. 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 elderlies as an important indicator of the level of autonomy they enjoy [103]. For example, if the level of ADLs is considered sufficient, it can indicate the slowing down of mental illness. If level of ADLs are considered insufficient, this could suggest that elderly people increase physical activity as an effective strategy to maintain independence [104]. If information about daily routine could be automatically collected, it could serve as a crucial reference for the prevention of NCDs and for prescription of behavioral therapies. Hence, we install a Kinect v2 sensor on the ceiling of an elderly subject to monitor the independent ability of the elderly person who is living independently. Meanwhile, an ADLs dataset is collected to test our proposed algorithms.

3.4.2.1 Sensor Installation

When other benchmarking datasets such as PKU-MMD [28] or NTU RGB+D 120 [36] 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 set-up, we collected a dataset that is small yet sufficient for developing skeleton-based HAR models that can be used for home healthcare. Figure 3.5 shows two examples of ADLs that we collected data for. We provide here 5 sampled RGB frames captured by the Kinect V2 sensor.

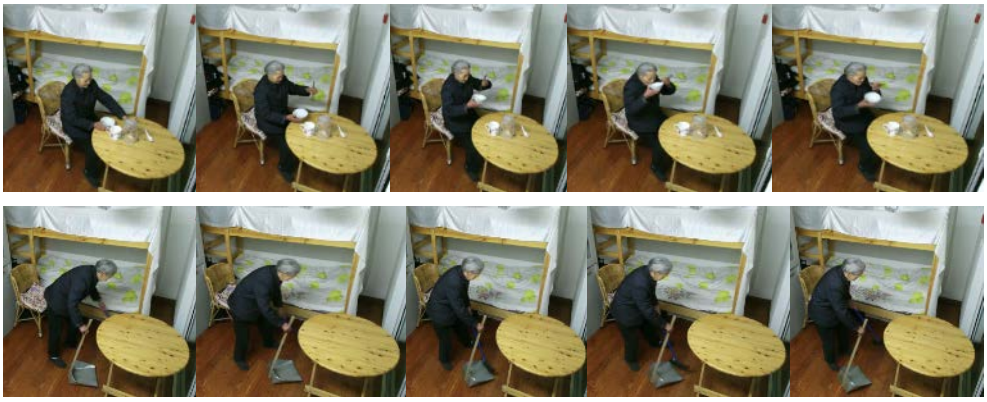


Figure 3.5 Sample activities of “eat with chopsticks” and “sweep the floor”

3.4.2.2 Collecting ADLs

Publicly available datasets like the NTU RGB-D 120 dataset [36] is an expanded version of the NTU RGB-D dataset [21] 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 will usually perform as in the morning routine (see Table 3.5). The dataset is distinctive with existing benchmark datasets in three aspects. First, this dataset 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 3.5 ADLs Dataset in the Morning Routine

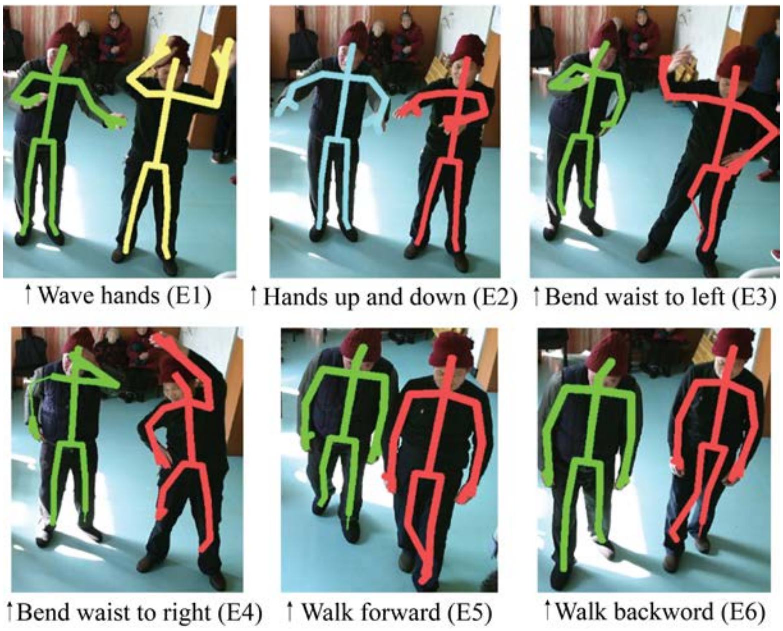
Activity Retrieved	Raw Data File Name	ADLs Type	Times
01 lie down	01_liedown_ getup	BADLs	9
02 get up	01_liedown_ getup	BADLs	9
03 comb hair	02_comb_hair	BADLs	11
04 pour water	03_pour_water_drink_water	BADLs	9
05 drink water	03_pour_water_drink_water	BADLs	9
06 eat with chopsticks	04_eat_with_chopsticks	BADLs	10
07 eat with iron spoon	05_eat_with_ironspoon	BADLs	10
08 eat with pottery spoon	06_eat_with_potteryspoon	BADLs	12
09 tidy table	07_tidy_table	IADLs	16
10 wipe table	08_wipe_table	IADLs	9
11 sweep the floor	09_sweep_the_floor	IADLs	19
12 wear shoes	10_wear_shoes	BADLs	17

3.4.3 Nursing Home Environment

Among the NCDs, it is reported that Alzheimer’s Diseases (AD) or other degenerative brain diseases among elderly people have been increasing more significantly when compare with other NCDs [105]. It is surveyed that the average duration of the Mild Cognitive Impairment (MCI) stage is seven years, which is a long-time dysfunction process where noticeable symptoms like decreased work performance and increased forgetfulness will appear gradually [106]. Early prevention and preparation at the MCI stage is essential for mitigate the significant deterioration of Quality of Life (QoL) for Alzheimer patients. Researchers are struggling to develop criteria for symptoms of Alzheimer at early stages to decelerate and even prevent the memory and cognitive decline progress. Existing MCI diagnosis methods could be categorized to core clinical criteria and research criteria or biomarkers [107]. Although biomarkers like cerebrospinal fluid, molecular neuroimaging with PET, and structural MRI analyses are objective, accurate, and universally useful, definite diagnosis of AD still half relies on clinical definitions of MCI. However, research criteria usually require at least six months of symptom appearance to make definite diagnosis. There is no empirical evidence that cognitive screening could provide effective support for decision making [108]. With such a background, we propose a vision-based HAE method in Chapter 6 that is able to automatically capture some clinical definitions of MCI to ease both the diagnosis and behavioral treatment of Alzheimer disease. In this section, we give introduction of the morning exercise data collection method and the collected dataset. The rational of evaluating physical exercises is further described in Section 6.1.

3.4.3.1 Exercises and Subjects

In the nursing home, elderly people will do daily morning exercise that is led by an exercise leader. The leader will demonstrate the physical exercise in front of a group of elderly people, and the elderly people will follow the demonstration to mimic the exercise. We do not define novel exercises for the people in the nursing home as it might be intrusive to their daily administration and also difficulty for both the exercise leader and the elderly people. To maintain the natural scenario in the nursing home, all the actions in this study are the same exercise that they do every day. As Figure 3.6 shows, there are six activities like wave hands, hands up and down, bend waits etc. being collected in our nursing home dataset.



↑Wave hands (E1) ↑Hands up and down (E2) ↑Bend waist to left (E3)

↑Bend waist to right (E4) ↑ Walk forward (E5) ↑Walk backword (E6)

Figure 3.6 Examples of the 6 morning exercises in the nursing home

3.4.3.2 Subjects and Repetitions

The demographic information of 25 subjects who took part in the morning exercise data collection is given in Table 3.6. The average age of them is 68.4 years with standard deviation of 10.82 years, which is unlike the existing datasets that are performed by young subjects. The relatively younger ones like S10, S16, S18, and S20 are staffs in the nursing home. 10 subjects of the dataset are diagnosed with varied severity of Alzheimer as indicated in the “AD” column of Table 3.6. The severity is labeled by a range of number from 0 to 10, where 0 represents no AD and 10 represents the last stage of AD. Care givers will feed AD medicine to all the Alzheimer subjects in the nursing home. Table 3.6 also shows the number of action repetitions retrieved from the raw Kinect v2 data. Due to the failure of skeleton detection, some of them might have 0 repetition.

Table 3.6 Demographic Information and Number of Repetitions of Each Action for the Subjects

Subject ID	Subject demographic Information						No. of repetitions					
	Age	Gender	Weight (kg)	Height (cm)	AD		E1	E2	E3	E4	E5	E6
S01	76	Male	70	168	0		7	5	4	9	2	3
S02	72	Male	66	180	7		7	5	6	6	3	3
S03	60	Male	53.5	172	8		6	5	6	8	3	3
S04	68	Male	60	160	0		6	5	5	6	2	2
S05	62	Male	70	165	5		0	5	8	6	2	2
S06	72	Female	51	157	10		7	5	8	7	1	1
S07	68	Female	54	158	0		7	7	6	1	4	3
S08	92	Male	55	165	0		8	6	6	3	3	3
S09	86	Female	55	163	0		11	5	6	4	2	2
S10	54	Male	60	162	10		8	5	6	7	2	3
S11	67	Male	85	185	5		9	6	10	9	3	3
S12	83	Male	65	170	6		9	5	12	9	3	2
S13	81	Female	48	151	4		9	6	13	0	3	3
S14	64	Male	65	172	8		7	6	6	5	3	2
S15	67	Male	70	170	6		8	6	6	6	3	2
S16	57	Female	69	156	0		8	6	7	6	4	4
S17	70	Male	94	182	0		8	6	14	3	4	4
S18	56	Female	63.5	158	0		7	6	12	4	4	3
S19	84	Male	60.5	175	0		8	5	15	2	4	4
S20	55	Female	55	160	0		8	6	13	13	4	4
S21	77	Female	47	161	0		8	7	14	15	3	3
S22	60	Female	58	163	0		11	7	13	15	3	3
S23	55	Female	57.5	162	0		9	7	0	12	3	3
S24	66	Female	65	163	0		11	6	0	15	3	3
S25	58	Female	66	161	0		11	6	0	15	3	3

Chapter 4

An HARELCARE Framework

In this chapter, the mathematical notation of the skeleton and RGB video data throughout this thesis will be given in Section 4.1. Then a HAR framework HARELCARE is introduced for ADLs recognition. This framework is unlike the existing ones that considers single sensor modality like video [10], or single algorithm group like [40]. It involves the available technologies could be considered in its steps for developing practical algorithms. For example, multiple algorithms could be attempted in the Step 3 for developing an effective HAR algorithm. Based on the HARELCARE framework, two different HAR algorithms are proposed in Section 4.3. The experimental results on the ADLs dataset is presented in Section 4.4 with discussion.

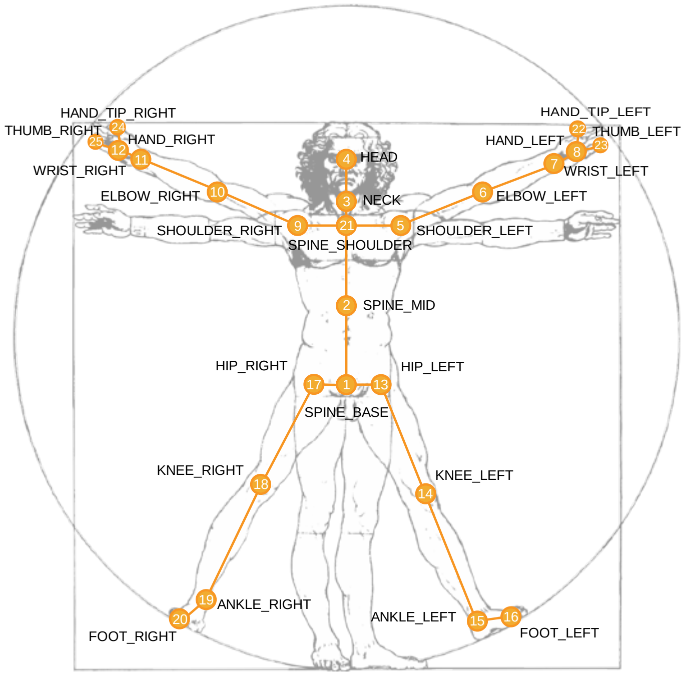


Figure 4.1 Skeleton joints of Kinect v2 sensor

4.1 A Mathematical Formulation of the Skeleton Data

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 Figure 4.1) 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 $\boldsymbol{j}_i^t$ has 8 attributes corresponding to its position and orientation. The position of joint i has 4 attribute features include 3-D cartesian coordinates of the position and its height from the floor, so that $\boldsymbol{j}_{i\_pos}^t = (j_{ix}^t, j_{iy}^t, j_{iz}^t, j_{ih}^t)$ with j_{ix}^t , j_{iy}^t, and j_{iz}^t correspond to the value of the x-, y- and z-coordinates, while j_{ih}^t indicates the vertical distance from the ground. The orientation of joint i is represented by a quaternion that has a set of values X, Y, Z, and W, which could be transferred to yaw, roll, and pitch values of the joint. So that $\boldsymbol{j}_{i\_ori}^t = (j_{iX}^t, j_{iY}^t, j_{iZ}^t, j_{iW}^t)$ with j_{iX}^t j_{iY}^t, j_{iZ}^t, and j_{iW}^t correspond to the values of X, Y, Z, and W. The i-th exercise repetition 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, $\boldsymbol{J}^{(i)} = [\boldsymbol{j}^1, \boldsymbol{j}^2, \dots, \boldsymbol{j}^t, \dots, \boldsymbol{j}^T]$. With a total of N samples from all exercise repetitions, the dataset could be represented as$\boldsymbol{J}^{(i)} = \{\boldsymbol{J}^{(i)} \mid i = 1,2, \dots, N\}$.

4.2 A Framework for ADLs Recognition

The raw data that we make use of for HAR are obtained from a Kinect sensor and can be represented, as discussed above, as $\boldsymbol{J}^{(i)} = \{\boldsymbol{J}^{(i)} \mid i = 1,2, \dots, N\}$. This set of raw data is therefore a set of multivariate, spatial-temporal data. Traditionally, algorithms like the DTW [37], HMM [38], and SVM [39] have been proposed for developing predictive models for HAR based on skeleton data. More recently, DL algorithms [109] 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.

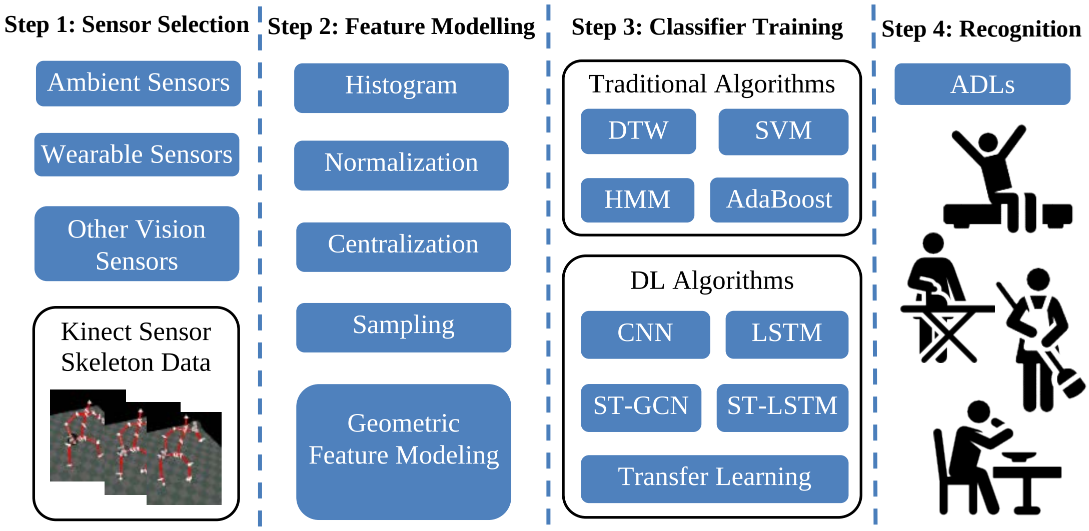


Figure 4.2 HARELCARE: ADLs recognition framework

Towards this goal, we propose a 4 steps framework as shown in Figure 4.2 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 might be able to deliver better recognition performance. For example, the use of the AdaBoost algorithm for the recognition of a single activity recognition could be implemented in the proposed framework [97]. Based on the use of the features transformed from the raw data, an 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 [110].

For higher accuracy, many DL-based algorithms 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 CNN or LSTM algorithms for training [86] [87]. 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 [88] [41]. 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 [62]. In addition to all these, there has recently been some effort to use multimodal approaches to HAR [111]. Instead of scaling up on the input data, spatial and temporal DL models like ST-LSTM [112], ST-GCN [113] have been shown to be quite effective in handling 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 the problem, we propose here a transfer learning method [114] that can be used for post-processing after a ST-GCN algorithm is used.

Transfer learning could be an option for agile HAR application development, however, to further improve the HAR accuracy, novel algorithms should be proposed. Beyond the framework, we originally proposed a vision-based multimodal HAR method that fuses the skeleton modality and RGB video modality at feature level.

4.3 HAR Algorithms

We propose two algorithms for practical HAR concerns when large dataset is not available or convenient to collect for training. The first algorithm is based on traditional feature extraction while the second one is based on transfer learning.

4.3.1 ABFE Algorithm

One first HAR algorithm that we can use for the proposed framework is the AdaBoost algorithm [115]. A typical AdaBoost algorithm can be trained to recognize a particular activity by developing a binary classifier. In other words, in the case that there are multiple activities to be recognized, we develop a binary classifier for each of them using the AdaBoost algorithm which we describe as follows.

4.3.1.1 Feature Modeling and Extraction

Given T skeleton joint frames $\boldsymbol{J}^{(i)} = [\boldsymbol{j}^{\boldsymbol{1}}, \boldsymbol{j}^{\boldsymbol{2}}, \dots, \boldsymbol{j}^{\boldsymbol{t}}, \dots, \boldsymbol{j}^{T}]$, for $\boldsymbol{j}^{t} = (\boldsymbol{j}_1^t, \dots, \boldsymbol{j}_i^t, \dots, \boldsymbol{j}_{25}^t)$ and $\boldsymbol{j}_{\boldsymbol{i}}^{\boldsymbol{t}} = (j_{ix}^t, j_{iy}^t, j_{iz}^t)$, it will be transformed to inter joints and intra joint features with transformation functions $\boldsymbol{f} = (f_1(\cdot), f_2(\cdot), \dots, f_i(\cdot), \dots, f_K(\cdot))$, where $f_i(\cdot)$ is one of K

such functions of $\boldsymbol{J}^{(i)}$ whose results can be considered inter- or intra-joint features $\tilde{\boldsymbol{f}} = (\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_i, \dots, \tilde{f}_K)$. Some representative functions that are used are given in Table 4.1 below. For example, the joint position distance is one of the inter joints features that could be calculated as

$$\tilde{f}_i = f_i(\boldsymbol{j}_k^t, \boldsymbol{j}_l^t) = \sqrt{(j_{kx}^t - j_{lx}^t)^2 + (j_{ky}^t - j_{ly}^t)^2 + (j_{kz}^t - j_{lz}^t)^2} \tag{4.1}$$

where $\boldsymbol{j}_k^t$ and $\boldsymbol{j}_l^t$ are two joints of the skeleton at time t. In the implementation, only representative joint pairs that considered as effective feature will be used.

Table 4.1 Examples of Feature Modeling Methods for AdaBoost

Inter joints features		**Intra joint features**
Joint position distance		Velocity in 3D space
Angles between 3 joints		Speed
Velocity of angles		Acceleration
Acceleration of angles		Muscle force

In other words, the training data $\boldsymbol{J}$ = {$\boldsymbol{J}^{(1)}$, $\boldsymbol{J}^{(2)}$, …, $\boldsymbol{J}^{(i)}$,… $\boldsymbol{J}^{(N)}$ } can be described in terms of these function as latent features $\tilde{\boldsymbol{f}} = (\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_i, \dots, \tilde{f}_K)$. The $\tilde{\boldsymbol{f}}$ is then used to train a binary classifier for each activity by using the AdaBoost algorithm [115]. Each feature $\tilde{f}_i$ will be used to build a weak classifier $g_i(\tilde{f}_i)$. In the scouting step of AdaBoost, with K weak classifiers, an expert pool which is represented as a matrix will be used to record the misses (with a 1) and hits (with a 0) of each classifier on every sample of the training set as shown in Table 4.2.

Table 4.2 Weak Threshold Classifiers Based on Features

	g_1	g_2	$\cdots$	g_k
$\boldsymbol{J}^{(1)}$	0	1	…	1
$\boldsymbol{J}^{(2)}$	0	0	…	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{J}^{(N)}$	0	0	…	0

Element in the expert pool matrix will be firstly initialized with weights $w_i = 1/N$. Then all the weights will be optimized by gradient descent. In the n-th iteration of the gradient update loop, the weight w_i^{n+1} will be updated by $\alpha_m = ln\,((1 - err_n)/err_n)$ as $w_i^{(n)} e^{\pm \alpha_n}$, where err_n is calculated by Equation (4.2).

$$err_n = \frac{\sum_{i=1}^{N} w_i e^{\alpha_n}}{\sum_{i=1}^{N} w_i} \qquad (4.2)$$

The final strong classifier $G(\boldsymbol{f})$ in (2) is a sign function of the sum of the top ten features selected from the expert pool.

$$G(\boldsymbol{f}) = sign\left(\sum_{m=1}^{K} \alpha_m \, g_m(f_m)\right) \qquad (4.3)$$

With M strong classifiers $G = \{G_1, G_2, \dots, G_M$ for M activities, the body frames $\boldsymbol{J}^{(i)}$ of each activity will be fed to the M strong classifiers G to generate a feature matrix $\boldsymbol{C}^{(i)} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_i, \dots, \boldsymbol{c}_M]$. Figure 4.3 shows visualized views of the feature matrix $\boldsymbol{C}^{(i)}$ and the vector $\boldsymbol{c}_i$ retrieved from all classifiers G and one specific classifier G_j, respectively.

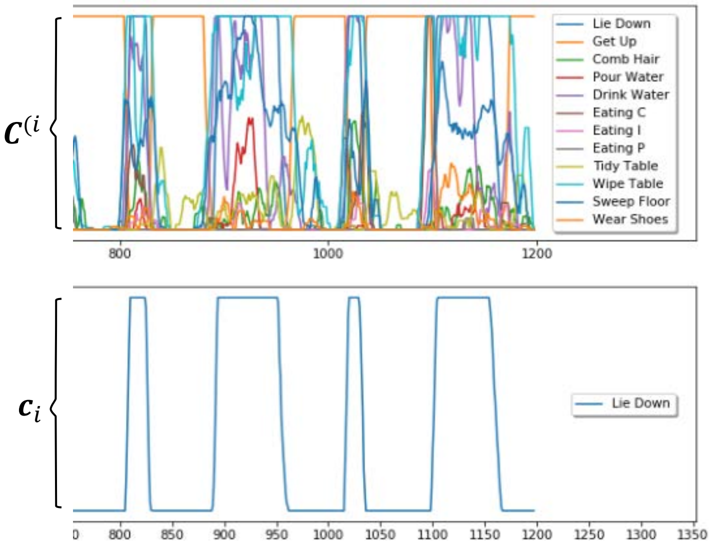


Figure 4.3. Visualization views of the low dimensional representation of an activity

We name this feature generation method as Adaptive Boosting Feature Extraction (ABFE) and summarized it in Algorithm 1, which is a feature-level method that reduces the dimension of the skeleton frames. The extracted feature matrix $\boldsymbol{C}^{(i)}$, for $\boldsymbol{c}_i \in \mathbb{R}^T$ could then be fed to different DL models to be trained for inferring its activity.

Algorithm 1: ABFE Algorithm

```
Data: $\mathbf{J} = \{\boldsymbol{J}^{(i)} \mid i = 1, \dots, N\}$ dataset for training
Result: $\boldsymbol{C} = \{\boldsymbol{C}^{(i)} \mid i = 1, \dots, N\}$ low dimensional representation of $\mathbf{J}$
$\boldsymbol{f} = (\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_i, \dots, \tilde{f}_K)$ = features transformed from $\boldsymbol{J}^{(i)}$
$g_i(\tilde{f}_i)$ = weak classifier in the AdaBoost expert pool
N = number of activities
for $n = 1$ to $M$
    $G_n(\boldsymbol{f})$ = strong binary classifier formed by the top 10 weak classifiers $\{g_i(\tilde{f}_i) | i = 1, \dots, 10\}$
end
return $G = \{G_1, G_2, \dots, G_M\}$
for $i = 1$ to $N$ do
    for $j = 1$ to $M$ do
        $\boldsymbol{c}_j$ = feature vector generated by $G_j$
    end
    $\boldsymbol{C}^{(i)} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_N]$ low dimensional representation of $\boldsymbol{J}^{(i)}$
end
return $\mathbf{C} = \{\boldsymbol{C}^{(i)} \mid i = 1, \dots, N\}$
```

4.3.1.2 Recognition Algorithm

Given the extracted low dimensional features $\boldsymbol{C}^{(i)} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_i, \dots, \boldsymbol{c}_M]$, $\boldsymbol{c}_i \in \mathbb{R}^T$ as depicted in Figure 4.3, we then use it to infer the activity by using an algorithm to represent the multivariate time series features. Specifically, we adopt the Multivariate Long Short Term Memory Fully Convolutional Network (MLSTM-FCN) algorithm proposed in [116]. The algorithm is built upon the long short-term memory (LSTM) RNNs that is capable to learn temporal dependencies. The LSTM modules is depicted by Graves [117] as

$$\mathbf{g}^c = \sigma(\mathbf{W}^c \mathbf{h}_{t-1} + \mathbf{I}^c \mathbf{c}_t)$$

$$\mathbf{g}^o = \sigma(\mathbf{W}^o\mathbf{h}_{t-1} + \mathbf{I}^o\mathbf{c}_t)$$
$$\mathbf{g}^f = \sigma(\mathbf{W}^f\mathbf{h}_{t-1} + \mathbf{I}^f\mathbf{c}_t)$$
$$\mathbf{g}^u = \sigma(\mathbf{W}^u\mathbf{h}_{t-1} + \mathbf{I}^u\mathbf{c}_t) \tag{4.4}$$
$$\mathbf{m}_t = \mathbf{g}^f \odot \mathbf{m}_{t-1} + \mathbf{g}^u \odot \mathbf{g}^c$$
$$\mathbf{h}_t = \tanh(\mathbf{g}^o \odot \mathbf{m}_t)$$

where $\mathbf{g}^c$, $\mathbf{g}^o$, $\mathbf{g}^f$, $\mathbf{g}^u$ are the activation vectors of cell state, output, forget and input gates, respectively. The recurrent weight matrices are denoted by $\mathbf{W}^c$, $\mathbf{W}^o$, $\mathbf{W}^f$ and $\mathbf{W}^u$. The projection matrices are represented as $\mathbf{I}^c$, $\mathbf{I}^o$, $\mathbf{I}^f$, $\mathbf{I}^u$. While $\mathbf{h}_t$ is the hidden state vector of the LSTM unit, σ is the logistic sigmoid function, and $\odot$ is the elementwise multiplication. On top of the LSTM unites, an attention mechanism that is a context vector depending on a sequence of annotations $(b_1, \dots, b_{T_c})$, where T_c is the maximum length of the sequence **c**. While the FCN module has a squeeze-and-excitation block that performs as will lead to the output for a single dimension as

$$\tilde{\boldsymbol{c}}_d = \mathrm{F}_{scale}(\mathbf{u}_d, \mathbf{s}_d) \tag{4.5}$$

where $\tilde{\mathbf{C}} = [\tilde{\boldsymbol{c}}_1, \dots, \tilde{\boldsymbol{c}}_M]$, $\mathbf{u}_d$ is the squeezed feature map generated by a channel-wise global average pooling, $\mathbf{s}_d$ is the excitation feature calculated from $\mathbf{u}_d$ by a sigmoid function followed by a ReLU function, and $\mathrm{F}_{scale}(\mathbf{u}_d, \mathbf{s}_d)$ denotes the channel wise multiplication of $\mathbf{u}_d$ and $\mathbf{s}_d$.

4.3.2 Transfer Learning

Collecting less data could ease the deployment of activity recognition, however, DL models are usually faced 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 has been proven helpful, which is a common strategy for transfer learning in the context of deep learning. [118] grouped transfer learning for HAR in three scenarios: inter-person, inter-device, and inter-ambiance. As the ST-GCN [113] shows the potential for representing spatial and temporal feature of skeleton data. We propose to use it as the backbone model and tune the trained weights of ST-GCN trained on the NTU-RGB+D dataset to our dataset to testify the effectiveness and efficiency of transfer learning. Our transfer learning method could be considered as inter-ambiance since we use different data collection environment with the 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 $\boldsymbol{\vartheta}_t = \{\boldsymbol{v}_t, \boldsymbol{\varepsilon}_t\}$, where $\boldsymbol{v}_t$ denotes the skeleton joints and $\boldsymbol{\varepsilon}_t$ demotes the skeleton bones time t, respectively. The neighbor set of a node v_{ti} is defined as $\mathbf{N}(v_{ti}) = \{v_{tj} | d(v_{ti}, v_{tj}) \leq D\}$, where D is the minimum path length of $d(\boldsymbol{v}_{ti}, \boldsymbol{v}_{tj})$. Suppose there are fixed number of K subsets in the $\mathbf{N}(v_{ti})$, every neighbor set will be labelled numerically with a mapping $l_{ti}: \mathbf{N}(v_{ti}) \rightarrow \{0, \dots, K-1\}$. Then the graph convolution could be computed as:

$$\mathrm{Y}_{\mathrm{output}}（\boldsymbol{v}_{ti}） = \sum_{\boldsymbol{v}_{tj} \in \mathbf{N}(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} \mathbf{X}(v_{tj}) \mathrm{W}(\mathrm{l}(v_{tj})) \tag{4.6}$$

where $\mathbf{X}(v_{tj})$ is the feature of $\boldsymbol{v}_{tj}$ that is equal to $(j_{jx}^t, j_{jy}^t, j_{jz}^t)$. While $\mathrm{W}(l(v_{tj}))$ is a weight function $\mathrm{W}(\mathrm{v}_{ti}, \mathrm{v}_{tj}): \mathbf{N}(v_{ti}) \rightarrow \mathrm{R}^{\mathrm{c}}$ that could be implemented by indexing a tensor of (c, K) dimension. The normalization term $Z_{ti}(v_{tj}) = |\{v_{tk} | l_{ti}(v_{tk}) = l_{ti}(v_{tj})|$ equals to the cardinality of the corresponding subset. With the specific partitioning strategy determined, the Equation (4.6) could be implemented with adjacency matrix A as

$$\mathrm{Y}_{\mathrm{output}} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathbf{\Lambda}_{\mathrm{k}}^{-\frac{1}{2}} \mathbf{A}_{\mathrm{k}} \mathbf{\Lambda}_{\mathrm{k}}^{-\frac{1}{2}} \mathbf{X} \mathrm{W}_{\mathrm{k}} \tag{4.7}$$

where $\boldsymbol{\Lambda}_{\boldsymbol{k}}^{\boldsymbol{ii}} = \sum_{\boldsymbol{j}} \mathbf{A}_{\boldsymbol{k}}^{\boldsymbol{ij}}$ is a degree matrix. Weiss et al. [114] performed a through survey for the transfer learning, which classifies transfer learning according to different categories as heterogeneous transfer learning solutions, homogeneous transfer learning solutions, and solutions addressing negative transfer that were further grouped to sub-categories. According to the categorization of [114], 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 2, which introduces the fine-tuning process.

Algorithm 2: Transfer Learning

```
    Data: $\mathcal{X}_T$, the input data J
    Weight: $W_S$, the model weight of the source domain; $W_T$ the model weight of the target domain.
    Output: $Y_{output}$, the output of the target domain
Load weights $W_S$ that is trained by using $\mathcal{X}_S$
Modify output layers of ST-GCN to adapt the output $\mathrm{Y}_{\mathrm{output}}$ of $\mathcal{D}_T$
Feed $\mathcal{X}_T$ to the modified model
For $i$ = 1 to epoch $M$ do
    For $j$ = 1 to Batch $N$ do
        Update $W_T$
    End
End
Use $W_T$ to infer $Y_{output}$
```

4.4 Results and Discussion

4.4.1 Evaluation Metric

To make the validation less biased than simply splitting the data to training set and test set, we adopt k-fold cross-validation evaluation method with *k* being set to 5 that follows the tradition and also makes the division of training and testing sets representative for measuring the fit of our model on the collected dataset. Cross-validation is popularly adopted for estimating the skill of a classification model when the sample size is limited like our small household dataset [119]. We use Top-1 accuracy as in Equation (4.8) that means the prediction must be the same as the label of the ground truth as the evaluation metric for classification tasks.

$$P = \frac{1}{N}\sum_{k=1}^{N} result_k = \begin{cases} 1 & if\ output_k^{top-1} = label_k \\ 0 & otherwise \end{cases} \tag{4.8}$$

With the accuracy measure set as Top-1, a confusion matrix, which is also known as error matrix, could be constructed as shown in Table 4.3. This matrix could be used to visualize the performance of a supervised classification algorithm with two or more classes [120]. In other words, the accuracy of an HAR prediction model can be gauged from the matrix. From the confusion matrix, we can derive the *precision* and *recall* measure of each class of activity. For better comparison of classification performance, a normalized confusion matrix could also be used.

In our experiments, for each confusion matrix corresponding to an activity, we accumulate 5 folds and normalize the entries in the confusion matrix for further comparison of the overall performances of different HAR algorithms.

Table 4.3 Confusion Matrix with N Classes

Actual Class	1	n_{11}	n_{12}	…	n_{1N}
	2	n_{21}	n_{22}	…	n_{2N}
	⋮	⋮	⋮		⋮
	N	n_{N1}	n_{N2}	…	n_{NN}
		1	2	…	N
		Predicted Class			

4.4.2 Results of Feature Extraction Method

4.4.2.1 Experimental Setting

We compare other two representative DL models mentioned before to justify the effectiveness of our proposed feature extraction step. The first model is ST-GCN [113] and the second model is ST-LSTM [112]. For ST-GCN, we keep the original implementation of the author and change the number of the output to 12. Since we use the Kinect v2, the joints number is set as 25. The number of people is 1 as our dataset only involves one subject. The maximum sequential frame length is set to 75 based on the statistics of the collected dataset. Hence, for one activity, the input to the ST-GCN in our experiment is a tensor with shape (3,25,75,1). Empirically, the initial learning rate is set to 0.1 and will decay to 1/10 of the precious learning rate at epochs of 40, 100, and 150. The predication result will be evaluated with an interval of 10 epochs. We train the model by setting the batch size to 64 and terminate the training at epoch 200 and show the best result throughout the training process. For ST-LSTM, we empirically follow the original hyper parameter setting except change the sequence length from 6 to 10, set the evaluation interval to 10, and terminate the training at epoch 200. The best result from all evaluations is selected to show in the experimental results.

In our method, the MLSTM-FCN model [116] that comprises of a fully convolutional block and a LSTM block that perform as feature extractors and finally concatenated together to a SoftMax layer is implemented by following the setting on the Arabic Voice dataset which is similar with the characteristic of our extracted features from the ABFE. We set the batch size and total epoch the same as the compared two DL models and show the best results in the next section.

By using the cross-validation method, both the skeleton dataset $\boldsymbol{J}$ and its transformed form as $\boldsymbol{C}$ are divided into five folds. Once one of the cross-validation folds is selected as testing set, the other three folds are used to train the model. There is no sample duplicated among the cross-validation folds. All the experiments are implemented on a Supermicro GPU Server (model SYS-7048GR-TR) with 4 GTX 1080 Ti GPUs. All the GPUs are used in each experiment.

4.4.2.2 Experimental Results

The experimental results for ST-GCN, LSTM and our method are listed in Table 4.4, which indicates the Top-1 accuracy of each cross-validation fold and their average accuracy. To investigate the training speed, we recorded the starting time and ending time, then calculated the differences of them. Table 4.5 gives the training time of all cross-validation folds. Specifically, for ST-GCN and ST-LSTM, the starting time and ending time are based on the generation time of the first checkpoint (at the epoch 10) and the last checkpoint (both at the epoch 200), respectively. For our method, we recorded the starting time and ending time of each training process.

Table 4.4. Experimental Results on Various Models (Accuracy in %)

Algorithm	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**
ST-LSTM	76.67	79.31	75.00	75.86	87.50	78.87
ST-GCN	66.67	77.41	67.86	55.17	75	68.42
Our Method	**92.86**	**92.59**	**96.15**	**96.15**	**96.43**	**94.84**

Table 4.5. Training Time of all Experimental Sets (Time in Seconds)

Algorithm	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**
ST-LSTM	0:24:12	0:25:26	0:28:24	0:24:29	0:30:35	0:26:37
ST-GCN	0:09:38	0:09:49	0:09:43	0:09:48	0:09:11	0:09:38
Our Method	**0:00:55**	**0:00:53**	**0:00:51**	**0:00:53**	**0:00:51**	**0:00:53**

Table 4.6 Accumulated and Normalized Confusion Matrix of ST-LSTM

True Label

1	0.78	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.67	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44
5	0.00	0.00	0.00	0.22	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.11
6	0.00	0.00	0.00	0.40	0.00	0.10	0.10	0.40	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.50	0.00	0.00	0.20	0.30	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.17	0.00	0.08	0.25	0.50	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.89	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.37	0.00	0.63	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	1	2	3	4	5	6	7	8	9	10	11	12

Predicted Label

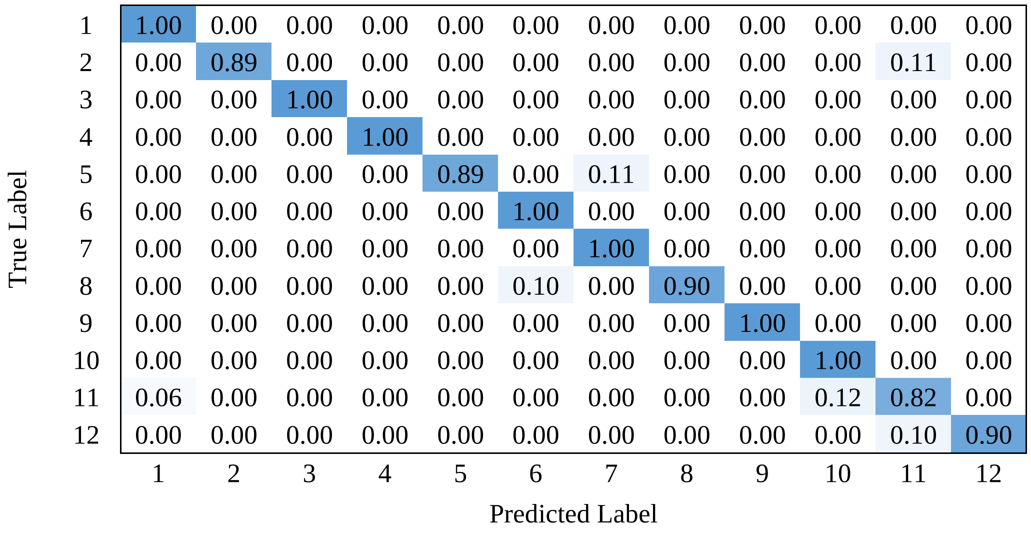
Table 4.7 Accumulated and Normalized Confusion Matrix of ST-GCN

True Label

1	0.33	0.33	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.11
2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.78	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.11	0.78	0.00	0.11	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.80	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.58	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	1	2	3	4	5	6	7	8	9	10	11	12

Predicted Label

Table 4.8 Accumulated and Normalized Confusion Matrix of Our Method



True Label	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.89	0.00	0.11	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.90	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
11	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.82	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.90

Predicted Label

Except showing the Top-1 accuracy and the training time, to further ease the comparison of the implemented methods, we show the accumulated and normalized confusion matrices of all three algorithms in Tables 4.6, 4.7, and 4.8.

4.4.2.3 Effectiveness Evaluation

According to the experimental results of Table 4.4, the extracted feature matrix from our proposed ABFE algorithm achieved the highest Top-1 accuracy in all five cross-validation folds. The average accuracy with the value of 94.84% is significantly better than that of data driven methods. From the results of ST-GCN and ST-LSTM in their cross-validation folds, we could observe that data driven methods requires the samples in the test set to have at least similar samples in the training sample to achieve the recognition or overfitting. It indicates that our feature extraction method successfully reduces the data size and maintains and even surpass the discriminative power of the skeleton data.

From the confusion matrix, we could have a closer look at which activities are failed and which activities are easy to be recognized by different algorithms. The figures in

Tables 4.6 and 4.7 show that activity 6 (eat with chopsticks) and activity 7 (eat with iron spoon) are the most confusing ones for the tested data driven models. While the discriminative feature of these two relatively fine-grained activities is successfully captured by our method as shown in Table 4.8 that indicates zero failure throughout the 5 cross-validation folds for both activity 5 and activity 6. The ST-GCN has the ability to learn spatial and temporal patterns with big dataset such as in [21] and [28]. However, when encountered with some confusing fine-grained challenging activities like pour water, drink water, and eat with pottery spoon, both spatial temporal DL models failed to learn effective feature in all its cross-validation folds.

We also observed from the confusion matrices that different models have their advantage to recognize specific activities. For example, ST-LSTM is the best in recognize activity 2 (get up), ST-GCN and ST-LSTM are both good at recognizing activity 12 (wear shoes), while our method is good at more activities like activity 1 (lie down), activity 4 (pour water), and activity 6 (eat with chopsticks). When the number of activities increases in this job, we could solve this issue by grouping activities according to different locations. In such a way, we could decrease the dimension of feature matrices from the feature generation step.

4.4.2.4 Effectiveness Evaluation

Our proposed method also achieved the shortest average training time (53 seconds for 200 epochs) as shown in Table 4.5. The training time is significantly less than that of ST-GCN and ST-LSTM that spend around 10 and 25 times more training time than our method, respectively. It indicates that the proposed feature extraction algorithm successfully reduces the data size to a feature matrix. This could be the reason why it has shorter training time than that of the raw skeleton-based methods, which makes the training process more efficient than the others and could be an advantage to speed up deployment of real-world solutions.

4.4.3 Results of Transfer Learning Method

4.4.3.1 Experimental Setting

For our experiments, we used k-fold cross-validation with k following the convention set to 5 [119]. 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.

4.4.3.2 Experimental Results

Table 4.9 provides the Top-1 accuracy of both the ST-GCN model and our transfer learning algorithm. From the figures 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.

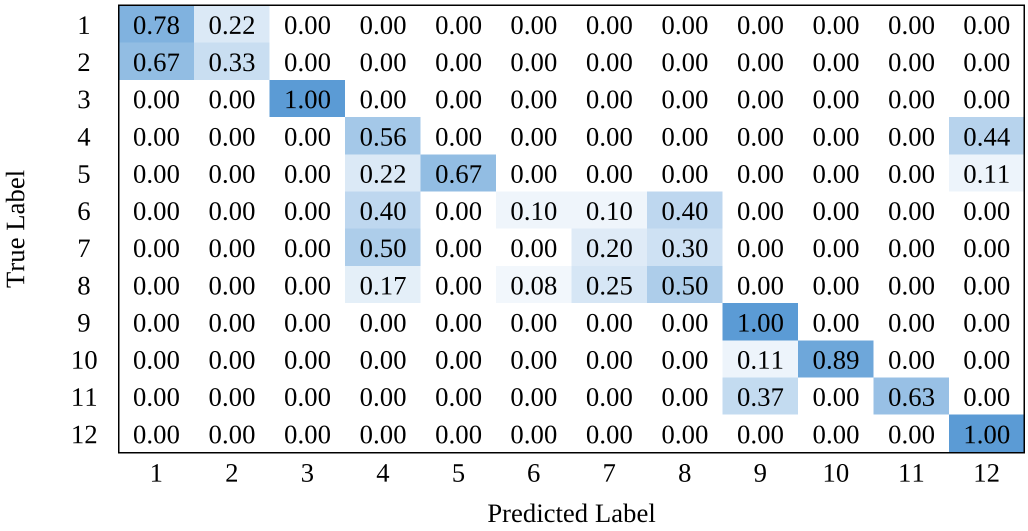
We also observe that some challenging fine-grained activities like eating with chopsticks or iron spoon are not exist in the source domain, but the similar activities like eating in the NTU-RGB+D should helped the improvement of the target domain (i.e.: our ADLs dataset). However, for activities 4 and 10 (i.e.: pour water and wipe table), the lack of similar activities in the source domain leads to the low recognition accuracy.

Table 4.9 Top-1 Accuracy on Three DL Models and Transfer Learning (Accuracy in %)

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

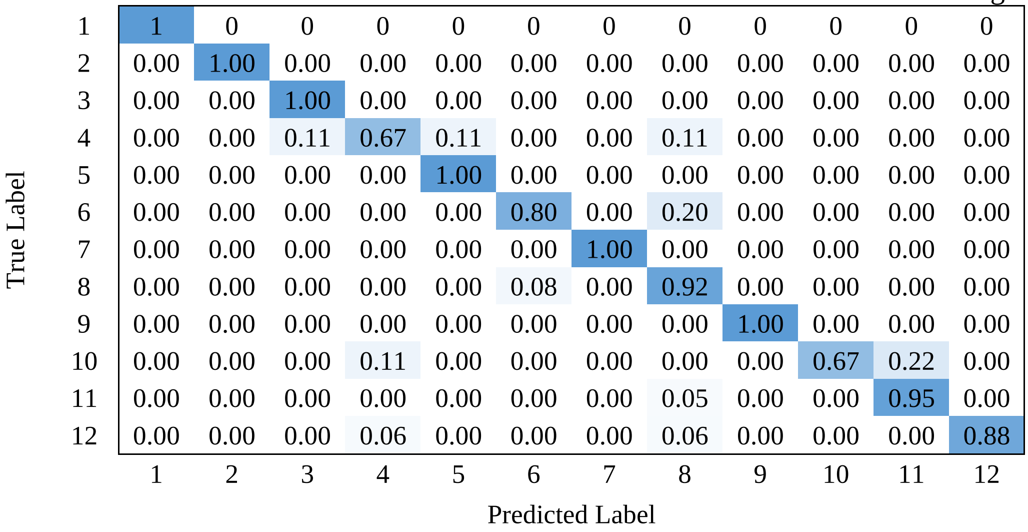
Other than showing improvement of the Top-1 accuracy by using transfer learning, the confusion matrices of them are visualized to further investigate the improvement as indicated in Table 4.10 and Table 4.11, respectively. It is noted that ST-GCN could not performed well on some activities like “get up”, “eat with chopsticks”, and “eat with iron spoon” (see Table 4.10). There is still some improvement space when tackling with a relatively small dataset that is tend to overfit 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 4.11, the recognition accuracy of the proposed transfer learning algorithm is close to optimal and there is little space for further improvement. Closer examination of the cases that the model failed to correctly recognize is due mainly to the big bias or unexpected noise like failure of skeleton detection by the Kinect v2 sensor.

Table 4.10 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	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.67	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44
5	0.00	0.00	0.00	0.22	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.11
6	0.00	0.00	0.00	0.40	0.00	0.10	0.10	0.40	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.50	0.00	0.00	0.20	0.30	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.17	0.00	0.08	0.25	0.50	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.89	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.37	0.00	0.63	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 4.11 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	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.11	0.67	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.80	0.00	0.20	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.92	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.67	0.22	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.95	0.00
12	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.88

Chapter 5

Multimodal HAR Method

We focus on two sets of features extracted from two data modalities (skeleton and RGB signals) as input of our multimodal HAR method. State-of-the-art spatial temporal GCNs like ST-GCN [44], AGC-LSTM [48], and DGN [49] could learn effective representation from the skeleton modality that has spatial importance for different skeleton joints. Meanwhile, models like I3D [56] and S3D [57] have the potential to learn discriminative features directly from video inputs but require huge computational resources. On the other hand, DL models like VGG nets [121] and ResNet [122] are effective to gain RGB features from images but are usually encountered with overfitting for datasets that are not big enough. Consequently, fusing the complementary skeleton and RGB features could be beneficial for action recognition.

There are various methods available for feature fusion. The fusion strategy relies on the characteristics of the involved data modalities. Most existing fusion is either at the decision level or at later layer concatenation, which is a lack of considering the borrowing of features from one modality to improve the performance of another modality, and eventually improves the performance. Since from the data of the RGB modality, it could be easy to make DL models with high representation power to overfit to the background noise. We propose a multimodal deep learning model called Model-based Multimodal Network (MMNet) (see Figure 5.1) that borrows spatial knowledge from the skeleton modality to alleviate this issue. On the other hand, the lack of appearance features from the skeleton modality renders it hard to distinguish the activities, especially those with object interactions. Hence, we claim that a proper feature representation of RGB modality could contribute back to the skeleton modality and boosts the ultimate performance.

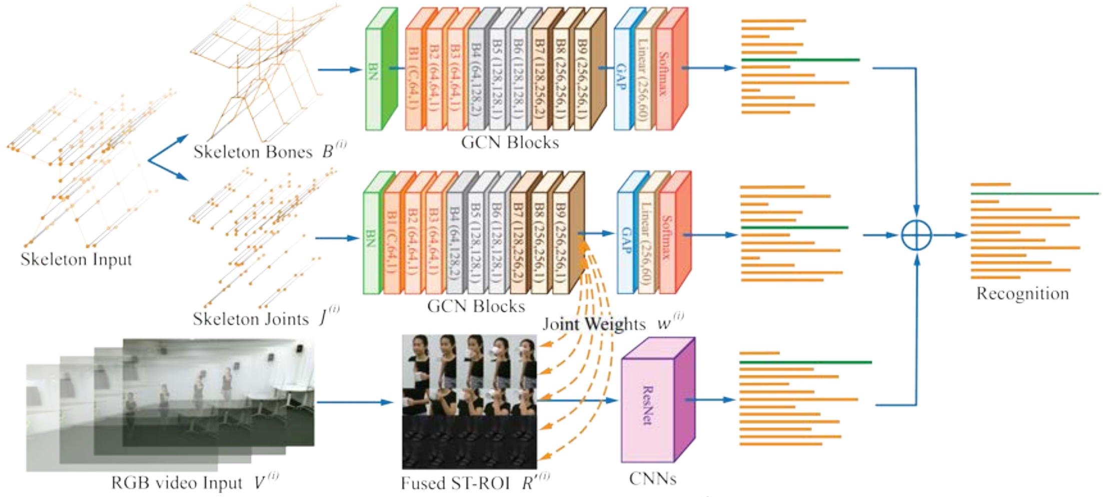


Figure 5.1 The architecture of the proposed MMNet. $w^{(i)}$ represents the spatial attention weights derived from the graph representation of the skeleton joints, which guides the focus of ST-ROI that is transformed from the RGB video input $V^{(i)}$. After this model-based data fusion, the skeleton focused ST-ROI $R'^{(i)}$ will be fed to the ResNet to generate modality specific prediction. The skeleton input will generate two separate predictions from its joints and bones, which is aggregated with the prediction of the RGB modality to deliver the ensemble recognition result.

5.1 Our MMNet Model

In this section, we introduce the proposed multimodal DL architecture by first describing the subnetworks utilized to learn features from the skeleton and RGB modalities and then elaborating feature fusion mechanisms between the two modalities.

We use the same notation for the skeleton modality with that of Section 4.1. For the RGB modality, let us notate $\boldsymbol{V} = \{\boldsymbol{V}^{(i)} \mid i = 1,2, \ldots, N\}$ as the RGB modality that has N video samples for training. Then an ordered video sequence of an activity in the time interval $[1, \mathrm{T}]$ could be represented as $\boldsymbol{V}^{(i)} = [\boldsymbol{f}_1^{(i)}, \dots, \boldsymbol{f}_t^{(i)}, \dots, \boldsymbol{f}_T^{(i)}]$, where $\boldsymbol{f}_t^{(i)}$ is the frame at time t.

5.1.1 Construct ST-ROI from RGB Modality

Intuitively, video-based models like I3D [56] and S3D [57] could be the first choice to learn discriminative features from the RGB modality. However, these models require huge computational resources of RAM and GPU memory and will take longer to converge. From our observation, even with pre-training, those models could not converge well with some datasets. Hence, we propose to build a spatial temporal ROI from the RGB modality and use general CNN models to retrieve effective features. Unlike the method proposed in [68] [69] [70] that focus on the appearance features of two hands, we build the spatial region of interest (ROI) that focuses on body parts including head, two hands and two feet in a temporal manner.

Let us notate $\boldsymbol{V} = \{\boldsymbol{V}^{(i)} \mid i = 1,2, \dots, N\}$ as the RGB modality that has N video samples for training. Then an ordered video sequence of an activity in the time interval $[1, \mathrm{T}]$ could be represented as $\boldsymbol{V}^{(i)} = [\boldsymbol{f}_1^{(i)}, \dots, \boldsymbol{f}_t^{(i)}, \dots, \boldsymbol{f}_T^{(i)}]$, where $\boldsymbol{f}_t^{(i)}$ is the frame at time t. To crop the spatial ROI from an activity video, we use joints of the skeleton retrieved with the OpenPose algorithm introduced in [123], which is relatively more accurate than the skeleton of the Kinect v2 sensor. Given an RGB frame $\boldsymbol{f}_t^{(i)}$, it could transformed to a spatial ROI with function $\mathrm{g}(\cdot)$. We define such a spatial transformation function as

$$R_{tj}^{(i)} = \mathrm{g}\left(\boldsymbol{f}_t^{(i)}, \boldsymbol{o}_{tj}^{(i)}\right), j \in (1, 2, \dots, K),\ K \leq M \tag{5.1}$$

where $R_{tj}^{(i)}$ and $\boldsymbol{o}_{tj}^{(i)}$ are the jth joint of the spatial ROI and the jth joint of the OpenPose skeleton at time t, respectively. K is the index of the skeleton joints, which is equal or smaller than the total number of the skeleton joints M . Given $\boldsymbol{V}^{(i)} = [\boldsymbol{f}_1^{(i)}, \dots, \boldsymbol{f}_t^{(i)}, \dots, \boldsymbol{f}_T^{(i)}]$, we then conduct a temporal sampling that selects L representative frames at time $\tau = \{\tau + interval \times l \mid l = 1, \dots, L, interval = T/L\}$ and concatenate them into a square ST-ROI as shown in the one subject case of Figure 5.2. For activities that have two subjects, we crop the ST-ROIs of both subjects as shown in the two subjects case of Figure 5.2. The ST-ROI significantly reduces the data volume of the RGB video modality and still preserves the object information of activities. The sub temporal ROI at time τ will have K sub spatial ROIs, which could be vertically concatenated and represented as $\boldsymbol{R}_\tau^{(i)}$. On the other hand, the sub spatial ROI of the jth joint will have L sub temporal ROIs and could be horizontally concatenated and represented as $\boldsymbol{R}_j^{(i)}$. The ST-ROI of $\boldsymbol{V}^{(i)}$ could then be notated as $\boldsymbol{R}^{(i)}$ that contains $K \times L$ sub ST-ROIs $\boldsymbol{R}_{\tau j}^{(i)}$.

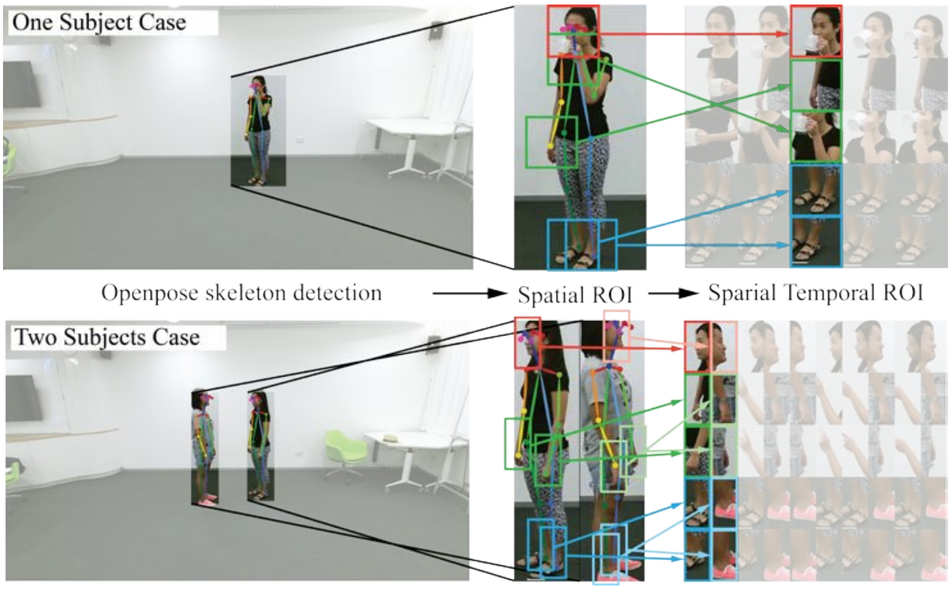


Figure 5.2 Process of constructing the spatial temporal ROI

5.1.2 Learn Joint Weights from the Skeleton Modality

For the skeleton modality, given a set of M joints in a skeleton frame observed at time t, let us represent it as $\boldsymbol{j}_t = (\boldsymbol{j}_{t1}, \dots, \boldsymbol{j}_{t2}, \dots, \boldsymbol{j}_{tM})$. The i-th training sample that starts at time $t = 1$ and ends at time T with skeleton frames collected at regular intervals can, therefore, be represented as a sequence of T skeleton frames, $\boldsymbol{J}^{(i)} = [\boldsymbol{j}_1, \boldsymbol{j}_2, \dots, \boldsymbol{j}_t, \dots, \boldsymbol{j}_T]$. With a total of N training samples, the skeleton modality and video modality of a dataset could be represented as $\boldsymbol{J} = \{\boldsymbol{J}^{(i)} \big| \, i = 1,2, \dots, N\}$. We adopt a spatiotemporal graph to model the spatial and temporal structure of $\boldsymbol{J}^{(i)}$. The structure of the Graph Convolutional Network (GCN) follows [44] and [124]. Figure 5.3 illustrates an example of the constructed spatial temporal graph of a skeleton sequence, where the skeleton joints are represented as vertices and the skeleton bones are represented as spatial edges (the orange lines in Figure 5.3a). For the temporal dimension, the corresponding skeleton joints between two consecutive skeleton frames are connected with temporal edges (the black lines in Figure 5.3a). The attribute of a vertex is the corresponding 3D coordinate values of each joint. The skeleton graph at time t could be symbolized as $\boldsymbol{\vartheta}_t = \{\boldsymbol{v}_t, \boldsymbol{\varepsilon}_t\}$, where $\boldsymbol{v}_t$ denotes the skeleton joints and $\boldsymbol{\varepsilon}_t$ denotes the skeleton bones, respectively. In this skeleton graph, the node set $\boldsymbol{v} = \{v_{tj} | v_{tj} = \boldsymbol{j}_{tj}, t = 1, \dots, T, j = 1, \dots, M\}$ contains all skeleton joints of a activity sequence. While the edge set $\boldsymbol{\varepsilon} = \{\varepsilon_t | \varepsilon_t = \boldsymbol{b}_t = (v_{ti} - v_{tj}), t = 1, \dots, T, i, j = 1, \dots, M\}$ represents skeleton bones of a skeleton activity sequence. With such a transformation, we will have the a sequence of bones from the skeleton modality and denote it as $\boldsymbol{B}^{(i)} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_t, \dots, \boldsymbol{b}_T]$.

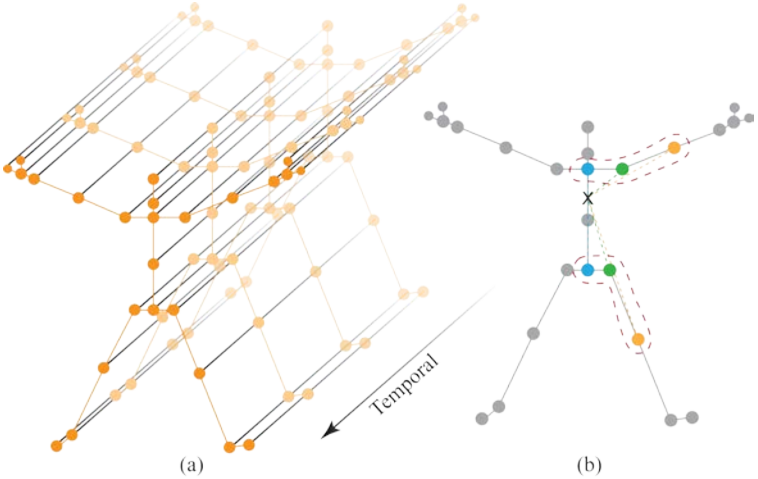


Figure 5.3 (a). The structure of a spatiotemporal graph. (b). The spatial mapping strategies. Different subsets are denoted with different colors. Green denotes the vertex itself; Yellow denotes the farther centrifugal subset; Blue denotes the closer centripetal subset

5.1.2.1 Graph Convolutional Operation

To represent the sampling area of convolutional operations, a neighbor set of a node υ_{ti} is defined as $\mathrm{N}(\upsilon_{ti}) = \{\upsilon_{tj} | d(\upsilon_{ti}, \upsilon_{tj}) \leq D\}$, where D is the minimum path length of $d(\upsilon_{ti}, \upsilon_{tj})$. The sketch in Figure 5.3b shows such a strategy, where × denotes the center of gravity of the skeleton. The sampling area $\mathrm{N}(\upsilon_{ti})$ is enclosed by the dot circled area. In detail, the strategy empirically uses 3 spatial subsets: the vertex itself (the green circle in Figure 5.3b); the centripetal subset that contains the neighboring vertices being closer to the center of gravity (the blue circle); and the centrifugal subset that contains the neighboring vertices being farther from the gravity center (the yellow circle). Suppose there is a fixed number of K subsets in the $\mathrm{N}(\upsilon_{ti})$, every neighbor set will be labelled numerically with a mapping $l_{ti}: \mathrm{N}(\upsilon_{ti}) \rightarrow \{0, \dots, K-1\}$. Temporally, the neighborhood concept is extended to temporally connected joints as $\mathrm{N}(\upsilon_{ti}) =$

$\{v_{qj} \middle| d(v_{tj}, v_{ti}) \le K, |q-t| \le \Gamma/2\}$, where Γ is the temporal kernel size that controls the temporal range of the neighbor set. Then the graph convolution could be computed as

$$\mathrm{Y_{output}}(\boldsymbol{v}_{ti}) = \sum_{\boldsymbol{v}_{tj} \in \mathbf{N}(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}^{t}(v_{tj}) \mathrm{W}(l(v_{tj})) \tag{5.2}$$

where $f_{in}^{t}(v_{tj})$ is the feature map that gets the attribute vector of $\boldsymbol{v}_{tj}$, $\mathrm{W}(l(v_{tj}))$ is a weight function $\mathrm{W}(\mathrm{v}_{ti}, \mathrm{v}_{tj})\colon \mathbf{N}(v_{ti}) \rightarrow \mathrm{R^c}$ that could be implemented by indexing a tensor of (c, K) dimension. $Z_{ti}(v_{tj}) = \left|\{v_{tk} | l_{ti}(v_{tk}) = l_{ti}(v_{tj})\}\right|$ is a normalization term that equals the cardinality of the corresponding subset.

5.1.2.2 Joint Weights

With implementation of graph convolution on the skeleton modality, the output of each vertex on the graph could be used to infer the importance of the corresponding skeleton joint. The feature map of the skeleton sequence could be represented as a tensor with size of (C, T, V), where V denotes the number of vertices, T denotes the temporal length and C denotes the number of attributes of the joint vertex. With the specific partitioning strategy determined, it could be represented by an adjacent matrix $\mathbf{A}$ with its elements indicating if a vertex $\boldsymbol{v}_{tj}$ belongs to a subset of $\mathbf{N}(v_{ti})$. The graph convolution is implemented by performing a $1 \times \Gamma$ classical 2D convolution and multiplies the resulting tensor with the normalized adjacency matrix $\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{A}\boldsymbol{\Lambda}^{-\frac{1}{2}}$ on the second dimension. With K partitioning strategies $\sum_{\mathrm{k=1}}^{\mathrm{K}} \mathbf{A}_{\mathrm{k}}$, Equation (5.2) could be transformed into

$$\mathrm{Y_{feature}} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \boldsymbol{\Lambda}_{\mathrm{k}}^{-\frac{1}{2}} \mathbf{A}_{\mathrm{k}} \boldsymbol{\Lambda}_{\mathrm{k}}^{-\frac{1}{2}} f_{\mathrm{in}} \mathrm{W_k} \odot \mathrm{M_k} \tag{5.3}$$

where $\boldsymbol{\Lambda}_k^{ii} = \sum_j(\mathbf{A}_k^{ij}) + \alpha$ is a diagonal matrix with α set to 0.001 to avoid values in a row being zero. $\mathrm{W_k}$ is a weight tensor of the 1×1 convolutional operation with $(C_{in}, C_{out}, 1, 1)$ dimensions, which represents the weighting function of Equation (5.2). $\mathrm{M_k}$ is an attention map with the same size of $\mathbf{A}_\mathrm{k}$, which indicates the importance of each vertex. $\odot$ denotes the element-wise product between two matrixes. $\mathrm{Y_{feature}}$ is a tensor with the size of (c, t, v) with c as the number of output channels, t as the temporal length and v as the number of vertices, which could be used to infer the activity class and transformed as joint weights to provide knowledge for the RGB modality. The joint weights that represent the importance of joints could be interpreted as

$$\mathrm{J_{weight}} = \frac{1}{tc}\sum_1^t\sum_1^c\sqrt{\mathrm{Y_{feature}}^2} \tag{5.4}$$

where t and c are the output dimensions of the convolutional graph that represents the temporal length and out channel, respectively. $\mathrm{J_{weight}}$ is a vector that contains weights for different skeleton joints.

5.1.3 Feature Fusion

We propose a spatial weight mechanism for the RGB frames to enable the machine being capable to focus on RGB features that will provide discriminative information. In an explainable way, this will make the machine more capable as it intuitively mimics activity recognition of human ways. Researchers also attempt to learn an attention weight from the RGB modality itself. From the result of [125] that has tested four variants of attention mechanism on the job of Convolutional LSTM [126], there are very few or even no performance improvements, although it decreases the model size.

Moreover, the gesture datasets used by [125] have very consistent backgrounds, which might make the attention mechanism even less effective for datasets that have varying complex backgrounds. Hence, we do not continue to explore the contribution of the attention mechanism in this job. Instead, we use the joint weights from the skeleton modality and multiply it with the ST-ROI to reduce the noise of the RGB modality. The weighted ST-ROI of the ith training sample $\boldsymbol{R'}^{(i)}$ could be mapped from $\boldsymbol{R}^{(i)}$. This process could be represented as a mapping function $h(\cdot)$ defined as

$$\boldsymbol{R'}^{(i)} = h\left(\boldsymbol{R}_j^{(i)}, J_{weight\_j}\right), j \in (1, 2, \dots, K) \tag{5.5}$$

where J_{weight_j} is the joint weight of the jth joint and $\boldsymbol{R}_j^{(i)}$ is the sub spatial ROI of the jth joint. Figure 5.4 shows the process of function (5) that denoises the RGB modality.

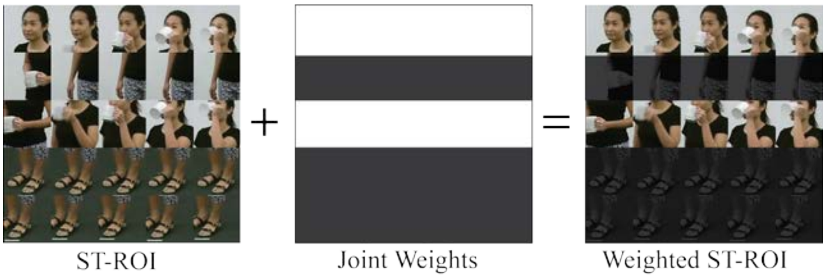


Figure 5.4 Multiply joint weights to the ST-ROI

5.1.4 Objective function

We build the end-to-end format of the multimodal algorithm with the sum of a collective loss items from different data modalities that are supervised by the activity label, which are explained below:

$$\mathcal{L} = \mathcal{L}_J\left(\hat{y}_J, y\right) + \mathcal{L}_B(\hat{y}_B, y) + \mathcal{L}_V(\hat{y}_V, y) \tag{5.6}$$

Skeleton joints modality loss -- $\mathcal{L}_J$

The skeleton joints input is fed into the graph convolution model introduced in Section 5.1.2. Hence the cross-entropy loss of skeleton joints could be defined as

$$\mathcal{L}_J\left(\hat{y}_{J^{(i)}}, y^{(i)}\right) = G_J(\Theta_J, J^{(i)}) - y^{(i)} = -\sum y^{(i)} log(S(\sigma(\hat{J}^{(i)}))) \tag{5.7}$$

where $\hat{J}^{(i)}$ represents the result of graph convolutional operation defined in Equation 5.3. σ denotes a fully connected layer that transforms the shape of $\hat{J}^{(i)}$ to a one hot representation. S is the Softmax function that transfer the recognition results to human understandable format.

RGB video modality loss -- $\mathcal{L}_V$

Recall that we have proposed the ST-ROI as the transformed form of the RGB video input, which will significantly reduce the data volume and maintain the core discriminative information for HAR. As the ST-ROI is intrinsically an 2D feature map, we adopt the ResNet proposed by He et al. [122]to learn features from it. The cross-entropy loss is typically adopted to optimize the model, which could be formulated as

$$\mathcal{L}_V\left(\hat{y}_{V^{(i)}}, y\right) = -\sum y^{(i)} log\left(G_V\left({R'}^{(i)}, \Theta_V\right) + {R'}^{(i)}\right) \tag{5.8}$$

where $G_V\left({R'}^{(i)}, \Theta_V\right)$ represents the residual mapping to be learned, Θ_V denotes the learnable weight that is based on the number of layers of the ResNet [122].

Skeleton bones modality loss -- $\mathcal{L}_B$

The skeleton bone modality is essentially a transformation of the skeleton joints modality, which proves more discriminative than the skeleton joints modality as in [124]. Hence, we also utilize the advantage of this transformed form of the skeleton joints modality. Recall that in the graph, the edge set is defined as $\boldsymbol{\varepsilon} = \{(v_{ti} - v_{tj}) | v_{ti}, v_{tj} = \boldsymbol{j}_{tj}, t = 1, \dots, T, i, j = 1, \dots, M\}$, which includes all the combination of the joint pairs represented in the adjacency matrix $\mathbf{A}$. We follow the transformation method in [124] and also base on the actual structure of the skeleton bones of the specific dataset. For example, given two joint vectors $v_{t1} = (x_1, y_1, z_1)$ and $v_{t2} = (x_2, y_2, z_2)$, then the bone vector could be calculated as $\boldsymbol{\varepsilon}_{t1} = v_{t1} - v_{t2} = (x_1 - x_2, y_1 - y_2, z_1 - z_2)$. We use the same graph convolutional operation to the skeleton bone modality, which could be formulated as

$$\mathcal{L}_B\left(\hat{y}_{B^{(i)}}, y^{(i)}\right) = G_B(\Theta_B, B^{(i)}) - y^{(i)} = -\sum y^{(i)} log(S(\sigma(\hat{B}^{(i)}))) \qquad (5.9)$$

where $\hat{B}^{(i)}$ denotes the output of Equation 5.3 fed with the skeleton bones $B^{(i)}$. σ and S are the same functions defined in Equation 5.7.

5.1.5 Training and optimization

Given the objective function, there are many deep learning multimodal data fusion strategies that could be adopted to pursue high recognition accuracy. For example, in [68], both the pose prediction loss that encourages the model reserve of the pose repression during training and the pose attraction loss that makes the skeleton attention more similar to humans can lead to attempting to improve performance. To ease the process of proving the effectiveness of our multimodal method, we avoid using such fine-tuning skills and hyperparameter tuning skills and adopt a vanilla training process to verify the effectiveness of our model as we have already proposed to fuse different data modalities at a feature level. More precisely, the training steps are illustrated in Algorithm 3.

```
Algorithm 3: Training of the proposed MMNet
Input: $\boldsymbol{V} = \{\boldsymbol{V}^{(i)} \mid i = 1,2,\dots,N\}$: RGB videos
       $\boldsymbol{J} = \{\boldsymbol{J}^{(i)} \mid i = 1,2,\dots,N\}$: skeleton joint coordinates
       $K$: the number of sub spatial ROIs
       $L$: the number of sub temporal ROIs
  1. Train GCN-Joints with skeleton modality of joint coordinates $\boldsymbol{J}$
  2. Construct a $K \times L$ ST-ROI $\boldsymbol{R}^{(i)}$ from the RGB video $\boldsymbol{V}$.
  3. Extract joint weights $J_{weight}$ by feeding $\boldsymbol{J}$ to the trained GCN-joints.
  4. Construct weighted ST-ROI $\boldsymbol{R}'^{(i)}$ from step 2 and 3.
  5. Train ResNet with $\boldsymbol{R}'^{(i)}$.
  6. Transform skeleton joint coordinates $\boldsymbol{J}$ to skeleton bone coordinates $\boldsymbol{B} =$
     $\{\boldsymbol{B}^{(i)} \mid i = 1,2,\dots,N\}$.
  7. Train GCN-bones with skeleton bones $\boldsymbol{B}$.
  8. Ensemble results from step 1, 5, and 7 to infer the activity class
Output: learned MMNet
```

5.2 Experiments and Results

In this section, we give an introduction of the benchmarking datasets selected in the experimental analysis and the comparison of the performance of our method with state-of-the-art methods.

5.2.1 Datasets

We conducted experiments on the two HAR datasets: NTU RGB+D Dataset [21], PKU-MMD [28] and Northwestern-UCLA Multiview dataset [29].

5.2.1.1 NTU RGB+D

NTU RGB+D dataset [21] was collected with by Kinect v2 sensors and contains over 56K samples of 60 different activities including individual activities, interactions between multiple people, and health-related events. The activities were performed by 40 subjects and recorded from 80 viewpoints. We followed the cross-subject and cross-view split protocol from [21]. Since this dataset provides multiple modalities of data from the Kinect v2 sensor, this dataset is highly suitable for testing multimodal HAR methods.

5.2.1.2 PKU-MMD

The PKU-MMD dataset [28] is another HAR dataset collected with Kinect v2. It contains 1076 long untrimmed video and skeleton sequences. The dataset is performed by 66 subjects in three camera views. With 51 activity categories annotated, we retrieved 21,545 valid activity sequences and 6 invalid samples that has no skeleton frames. Similar to NTU RGB+D, we adopt the two evaluation protocols (i.e., cross-subject and cross-view) recommended in [28]. For activity samples that have longer than 300 frames, we evenly select 300 frames form the samples.

5.2.1.3 Northwestern-UCLA Multiview

The Northwestern-UCLA Multiview dataset was collected by [29], which contains more interactions between human subjects and objects. The dataset has 12 action categories with each of them performed by 10 actors. It has 1,494 samples in total, which includes 518 samples from view 1, another 509 samples from view 2 and 467 samples from view 3. We follow the evaluation method in [68].

5.2.2 Implementation Details

For the RGB modality, the height and width of sub ST-ROIs of NTU-RGB+D, PKU-MMD and Northwestern-UCLA are 96, 96 and 48 pixels, respectively. We set both K and L to 5 to construct the ST-ROI. Therefore, the input size for NTU-RGB+D, PKU-MMD and Northwestern-UCLA datasets are 480×480, 480×480 and 240×240, respectively. The ST-ROI of the three datasets were resized to 225×225 and normalized before feeding them into ResNet. As the data volume of Northwestern-UCLA is relatively small, we perform random selection to the RGB video frames and randomly flip them. We adopted ResNet18 that has 18 layers for all the datasets. For NTU-RGB+D and PKU-MMD, we evenly selected the frames based on the video length for testing. For the skeleton modality, we follow the experimental setting of [124] for NTU-RGB+D while using the setting of [44] for PKU-MMD and Northwestern-UCLA. The SGD optimizer is utilized for all implementations with the initial learning rate set as 0.1 which is divided by 10 at the 10th and 50th epochs. The training process is terminated at the 80th epoch. The minibatch size is set to 64. All experiments are conducted on a workstation with 4 GTX 1080 Ti GPUs.

5.2.3 Results

5.2.3.1 Ablation Study

Tables 5.1, 5.2 and 5.3 show several experiments with different data modalities and their ensembled results. The results show considerable improvements by aggregating results of the RGB modality to the results of the skeleton modality with different training methods numbered as 4, 5, and 6 on NTU-RGB+D, PKU-MMD and Northwestern-UCLA Multiview. By comparing training methods 4 and 5, we could observe that the proposed joint weights mechanism is able to effectively improve the discriminative power of the ST-ROI features. Contrasting training method 5 with that of 6, tuning the weights of GCN together with the ResNet, could be of benefit for the RGB modality. However, in terms of the overall performance improvement when aggregating the results of the RGB modality and the skeleton modality, fixing the GCN weights by set it at evaluation mode will achieve better ensemble results for all three datasets. In Figure 5.5, we illustrated the effectiveness of the weighted ST-ROI method that improves the recognition accuracy of every activity of Northwestern-UCLA Dataset. It also indicates that even for some activities the accuracy of the RGB modality is not as good as the skeleton modality. However, it will still contribute to the overall performance. In Figure 5.6, we visualize the skeleton focused and normalized ST-ROI during training processes, which indicates that the irrelevant body parts from the RGB modality are masked by the learnt joint weights.

Table 5.1 Ablation Study for NTU RGB+D with Cross-Subject (CS) and Cross-View (CV) Protocol. ○ Means in Evaluation Mode. √ Means in Training Mode

#	Methods	$\mathcal{L}_J$	$\mathcal{L}_B$	$\mathcal{L}_R$	CS	CV	Avg
1	GCN-Joints	√	-	-	83.4	90.0	86.7
2	GCN-Bones	-	√	-	87.2	93.1	90.2
3	Ensemble (1+2)	○	○	-	88.3	94.5	91.4
4	ResNet18	-	-	√	76.4	80.5	78.5
5	ResNet18+Weights	√	-	√	86.8	89.3	88.1
6	ResNet18+Weights	○	-	√	82.8	87.4	85.1
7	Ensemble (3+4)	○	○	○	90.1	94.5	92.3
8	Ensemble (3+5)	○	○	○	90.2	96.8	92.4
9	Ensemble (3+6)	○	○	○	**92.5**	**97.3**	**94.9**

Table 5.2 Ablation study for PKU-MMD with Cross-Subject (CS) and Cross-View (CV) protocols. ○ means in evaluation mode. √ means in tuning mode

#	Methods	$\mathcal{L}_J$	$\mathcal{L}_B$	$\mathcal{L}_V$	CS	CV	Avg
1	GCN-Joints	√	-	-	91.5	92.4	91.2
2	GCN-Bones	-	√	-	93.4	95.1	94.3
3	Ensemble (1+2)	○	○	-	94.6	96.3	95.1
4	ResNet18	-	-	√	81.3	77.4	79.4
5	ResNet18+Weights	√	-	√	81.6	76.2	78.9
6	ResNet18+Weights	○	-	√	81.1	76.0	78.6
7	Ensemble (3+4)	○	○	○	95.8	97.1	96.5
8	Ensemble (3+5)	○	○	○	95.9	97.2	96.5
9	Ensemble (3+6)	○	○	○	**96.2**	**97.3**	**96.7**

Table 5.3 Ablation Study for the Northwestern-UCLA Multiview Action 3D Dataset with Cross-View Setting (Accuracy as a Percent). ○ Means in Evaluation Mode

#	**Methods**	$\mathcal{L}_J$	$\mathcal{L}_B$	$\mathcal{L}_R$	$V^3_{1,2}$
1	GCN-Joints	√	-	-	82.9
2	GCN-Bones	-	√	-	86.0
3	GCN-2s (1+2)	√	√	-	90.1
4	ResNet18	-	-	√	77.1
5	ResNet18+Weights	√	-	√	84.9
6	ResNet18+Weights	○	-	√	83.2
7	Ensemble (3+4)	○	○	○	90.5
8	Ensemble (3+5)	○	○	○	90.3
9	Ensemble (3+6)	○	○	○	**94.2**

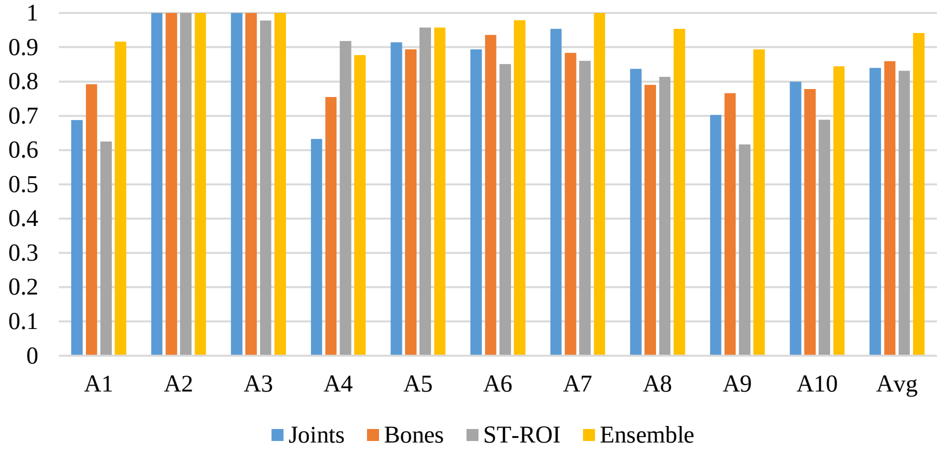
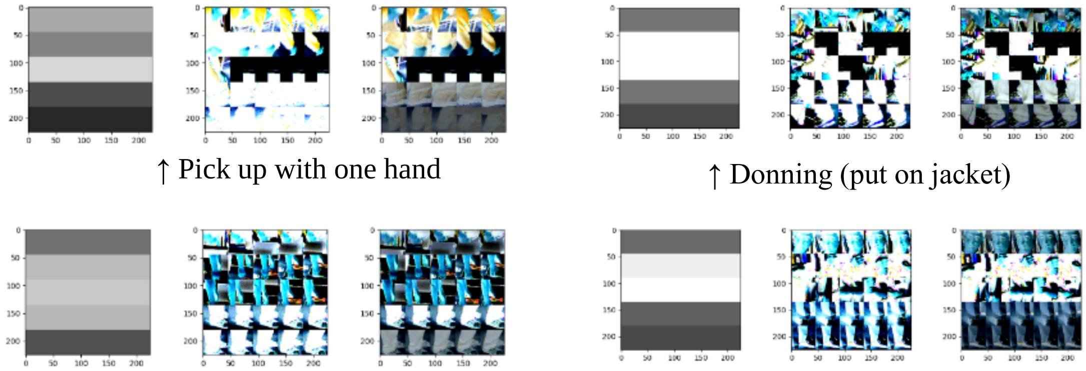


Figure 5.5 Recognition accuracy improvement in every activity of Northwestern-UCLA Dataset



↑ Pick up with one hand

↑ Donning (put on jacket)

↑ Pick up with two hands

↑ Doffing (take off jacket)

Figure 5.6 Visualization of joints weights, ST-ROI and joint weight ST-ROI (shown as in left, middle, and right of subplots, respectively) for different activities

5.2.3.2 Comparison with the State-of-the-Art

We show the performance comparison with the other state-of-the-art methods that use the skeleton modal, RGB modal and multimodal in Tables 5.4, 5.5 and 5.6 for the NTU-RGB+D, PKU-MMD and Northwestern-UCLA datasets, respectively. Our skeleton joint-weighted ST-ROI method achieved state-of-the-art performance on both datasets with a vanilla implementation by using the basic ResNet model which is ResNet18. Whereas, to achieve good performance, existing RGB based methods usually utilize much more complex CNN models like ResNet50 in [68].

Table 5.4 Results for NTU RGB+D with CS and CV evaluation settings

Methods	**Pose**	**RGB**	**CS**	**CV**	**Avg**
Lie Group [127]	√	-	50.1	52.8	51.5
Dynamic Skeletons [128]	√	-	60.2	65.2	62.7
Part-aware LSTM [21]	√	-	62.9	70.3	66.6
GCA-LSTM [129]	√	-	74.4	82.8	78.6
View-invariant [41]	√	-	80.0	87.2	83.6
ST-GCN [44]	√	-	81.5	88.3	84.9
DPRL+GCNN [130]	√	-	83.5	89.8	86.7
2S-AGCN [124]	√	-	88.5	95.1	91.8
AGC-LSTM [48]	√	-	89.2	95.0	92.1
DGNN [49]	√	-	89.9	96.1	93.0
C3D [131]	-	√	63.5	70.3	66.9
Glimpse Clouds [68]	-	√	86.6	93.2	89.9
DSSCA - SSLM [63]	√	√	74.9	-	-
STA-Hands [69]	√	√	82.5	88.6	85.6
Hands Attention [70]	√	√	84.8	90.6	87.7
Our MMNet	√	√	**92.5**	**97.3**	**94.9**

Table 5.5 Results for PKU-MMD with CS and CV evaluation settings

Methods	**Skeleton**	**RGB**	**CS**	**CV**	**Avg**
JCRRNN [132]	√	-	32.5	53.3	42.90
Skeleton boxes [133]	√	-	54.8	94.2	74.50
CNN-based [134]	√	-	90.4	93.7	92.05
STA-LSTM [135]	√	-	86.9	92.6	89.75
HCN [46]	√	-	92.6	94.2	93.40
SRNet [136]	√	-	93.1	97.0	95.05
Our MMNet	√	√	**96.2**	**97.3**	**96.75**

Table 5.6 Results for the Northwestern-UCLA Multiview Action 3D dataset with Cross-View Setting

Methods	Pose	RGB	$V_{1,2}^{3}$
Lie Group [127]	√	-	74.2
HBRNN-L [137]	√	-	78.5
View-invariant [41]	√	-	86.1
Ensemble TS-LSTM [138]	√	-	89.2
Hankelets [139]	-	√	45.2
nCTE [140]	-	√	68.6
NKTM [141]	-	√	75.8
Glimpse Clouds [68]	-	√	90.1
Our MMNet	√	√	**94.9**

5.2.3.3 Results on our ADLs Dataset

Table 5.7 Ablation Study and Comparison with Previous Methods on Our ADLs Dataset

Algorithm	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
ST-LSTM	76.67	79.31	75.00	75.86	87.50	78.87
ST-GCN	66.67	77.41	67.86	55.17	75	68.42
Feature Extraction	92.86	92.59	96.15	96.15	96.43	**94.84**
Transfer Learning (Joint)	83.33	83.10	92.86	93.10	95.83	89.64
ST-GCN Bone (Bone)	60.00	75.86	74.13	68.97	83.33	72.46
ST-ROI	30.00	24.14	42.86	46.88	58.33	40.44
ST-ROI + Joint	73.33	86.21	85.71	96.65	91.67	86.71
ST-ROI + Joint + bone (MMNet)	80.00	93.10	92.86	96.65	95.83	91.69

Table 5.7 shows the results of the multimodal method on our ADLs dataset with ablation study and comparison with four previous methods. We could obverse that the ST-ROI feature could not converge on the ADLs dataset, but it still not affects the performance of the skeleton modality when ensemble them together. In some Cross-Validation folds like fold 2 and fold 4, the performance of the multimodal method is even better than the feature extraction method. The low accuracy of the RGB modality is due to the insufficiency of training data in the ADLs dataset as it is collected for practical concerns and such a data-driven DL method is competitive but the traditional feature extraction is more effective in such a practical case.

5.3 Discussion

Multimodal HAR methods need to not only tackle with fusion of heterogeneous data modalities, but also handle the optimization problem. Unlike existing multimodal methods that usually focus on homogeneous data modalities like the skeleton joint and bone modalities, or RGB and optical flow modalities, our MMNet model focuses on heterogeneous data modalities. The novelty of our fusion scheme is fusing at the feature level, which is unlike existing methods that simple concatenate the feature at the representation level or decision level. Hence, we call it model-based fusion multimodal fusion based on the multimodal learning categorization in [22] and [142]. For optimization, we find there is a controversial decision between modality specific accuracy and ensemble accuracy as shown in the results of ablation study in Section 5.2.3.1. A model could pursuit either high modality specific accuracy or good ensemble accuracy. This is due to the tuning of the whole model will make the loss propagate back to the modality specific models and renders the representation share the features of various data modalities, which will decrease the contribution of modality specific features. To improve the ensemble accuracy, modality specific models should maintain their independent features to make use of the advantage of mutual complementary information between different data modalities.

The proposed MMNet relies on the RGB modality, hence, the RGB modality could not contribute the accuracy during dark light condition. But the skeleton modality of the solution could still work under poor light condition. Proper adapting mechanisms need to be developed to handle the change of light condition. Meanwhile, for real-time monitoring, online monitoring algorithms also need to be developed.

Chapter 6

Morning Exercise Evaluation for the Elderly in a Nursing Home

In this chapter, the rationale behind morning exercise evaluation for Alzheimer subjects in a nursing home is first provided. Then, an HAE method that is extended from the HAR framework is introduced.

6.1 Why Physical Exercise Evaluation?

It is surveyed that the average duration of the MCI stage is around seven years, which is a long-time deterioration process [106]. Early prevention and preparation at this stage is essential for alleviate significant decline of Quality of Life (QoL) for Alzheimer patients. Researchers are struggling to develop criteria for symptoms of Alzheimer's disease at early stages to decelerate and even prevent the memory and cognitive decline progress. Existing MCI diagnosis methods could be categorized to core clinical criteria and research criteria or biomarkers [107].

Clinical criteria refer to symptom-based methods or cognitive testing like Mini Mental State Examination (MMSE), Modified Mini-Mental State Examination (3MS) and Abbreviated Mental Test Score (AMTS), which is traditionally accepted as of reliable AD detection. However, it requires at least six months of symptom appearance to make definite diagnosis. There is few empirical results that verify the cognitive screening could improve decision making [108]. Human observation could also be performed by an informant to fill a questionnaire based on the daily cognitive functioning of a patient. Besides, more detailed description could be assessed by numeric scales like Clinical Dementia Rating (CDR), Global Deterioration Scale for Assessment of Primary Degenerative Dementia (GDS or Reisberg Scale), and Functional Assessment Staging Test (FAST) [143]. Since the emerging of numerous biomarkers, these clinical criteria or definitions for MCI are incorporated with research criteria to make probabilistic diagnosis of AD [144].

Research criteria includes biomarkers like molecular neuroimaging with PET, structural MRI, and cerebrospinal fluid analyses, which are used in the revised version of criteria named the National Institute of Neurological Disorders and Stroke–Alzheimer Disease and Related Disorders (NINCDS–ADRDA) [145]. Although biomarkers are objective, accurate, and universally useful, definite diagnosis of AD still half relies on clinical definitions of MCI. In the diagnostic criteria for AD proposed by [145], core diagnostic criteria include episodic memory impairment and cognitive changes. The progressive episodic memory impairment can be reported by patients or informants over more than 6 months, or objectively verified by memory testing. Most cases of cognitive changes are associated with the episodic memory disorder, which involve the domains like executive functions (EFs), language, praxis, complex visual processing and gnosis. The EFs could be impaired due to lack of sleep, stress, lack of exercise or loneliness. Excellent results have been achieved by much work on improving EFs in the elderly by conducting systematic physical training [146].

Research is accumulating to suggest that systematic exercise training increase strength, balance, and flexibility, and improve cardiovascular function, and in the meantime prevent cognitive dysfunction. For example, 11 patients with AD patients are benefited from a exercise program in a uncontrolled study in a hospital [147]. For elderly people without AD, randomized controlled trials have also indicated that physical exercise can reduce depression [148]. Linda et al. [148] investigated the seldom known deleterious effects of AD on body functional condition and proved that systematic exercise training can significant improve physical and affective measurements for patients with AD. Human and animal studies indicate that exercise targets many parts of brain function, and it has broad benefits to elderly populations on their overall brain health, brain plasticity and function, and improves resistance to neurodegenerative diseases [149]. The importance of aerobic exercise for maintaining neurocognitive performance is also verified by a meta-analytic review of randomized controlled trials [150]. Another meta-analysis also suggest that exercise can provide additional benefits for patients with AD [151]. A meta-analysis is a statistical analysis that investigates the results from multiple scientific studies like systematic review, and several clinical trials as shown in Figure 6.1, which aims to obtain a better knowledge of how well a treatment could work [152].

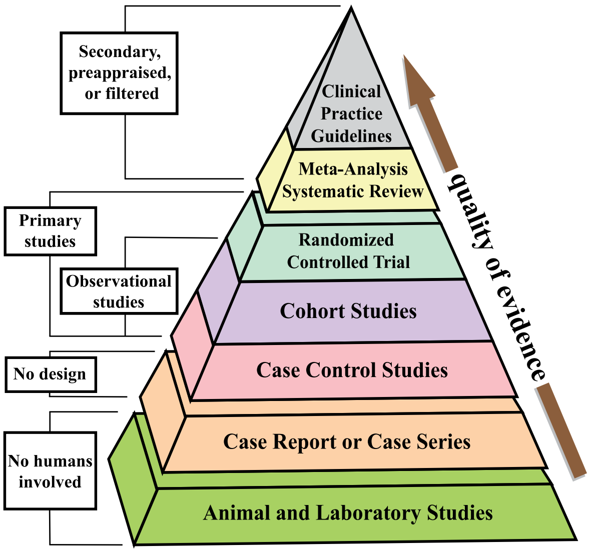

Secondary, preappraised, or filtered
Primary studies
Observational studies
No design
No humans involved
Clinical Practice Guidelines
Meta-Analysis Systematic Review
Randomized Controlled Trial
Cohort Studies
Case Control Studies
Case Report or Case Series
Animal and Laboratory Studies
quality of evidence

Figure 6.1 Hierarchy of evidence

In this thesis, we focus on the praxis domain of cognitive change and explore a vision-based method that is able to automatically capture some clinical definitions of MCI to ease both the diagnosis and behavioral treatment of AD. Praxis impairment could be impaired imitation, production, or recognition of gesture. Given the validated importance and effectiveness of systematic exercise training for preventing AD, we use a vision sensor to monitor the regular morning exercise in a nursing home. The motion data of the elderly's morning exercise is then collected and analyzed to infer their praxis health condition, which could potentially be used for supporting Alzheimer diagnosis and exercise-based therapy.

Hence, we propose a method called Two-Task Graph Convolutional Network (2T-GCN) to tackle the tasks. For supporting diagnosis, the proposed 2T-GCN will classify if the exercise has AD symptoms or not. While for supporting exercise-based therapy, our 2T-GCN could generate a numerical exercise quality evaluation score that reflects the praxis condition. Both two tasks are theoretically supported by the effectiveness and benefits of systematic exercise training. Meanwhile, we also explore machine learning methods that deliver consistent results with clinically verified evaluation results, which will be introduced in Section 6.3.4.2.

6.2 Our 2T-GCN Model

As we use the same Kinect v2 sensor to collect the exercise motion of the elderly in the nursing as introduced in Section 3.4.3, the raw skeleton data in one frame is always streamed as an ordered sequence of vectors. Each vector represents the position and orientation attributes of the corresponding human joint. A complete exercise repetition contains multiple frames with varied lengths for different repetitions. We adopt a spatiotemporal graph convolutional network to represent the structured information among these joints along both the spatial and temporal dimensions.

6.2.1 Graph Convolutional Network

6.2.1.1 Graph Construction

The construction of the skeleton graph follows the structure of ST-GCN [113]. Figure 6.2 illustrates an example of the structure of the spatial temporal skeleton graph, where the vertexes represents the skeleton joints and the spatial edges represents the skeleton bones which is natural connections of skeleton joints (the orange lines in Figure 6.2a). For the temporal dimension, the corresponding joints between two consecutives frames are represented by the connections of temporal edges (the black lines in Figure 6.2a). The position and orientation features of each joint as introduce in Section 4.1 are set as the attribute of the corresponding vertex. The skeleton graph at time t could be 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, respectively. In this graph, the node set $\boldsymbol{\upsilon}_t = \{\upsilon_{ti} | \upsilon_{ti} = \boldsymbol{j}_i^t, t = 1, \dots, T, i = 1, \dots, 25\}$ contains all joints in the skeleton sequence.

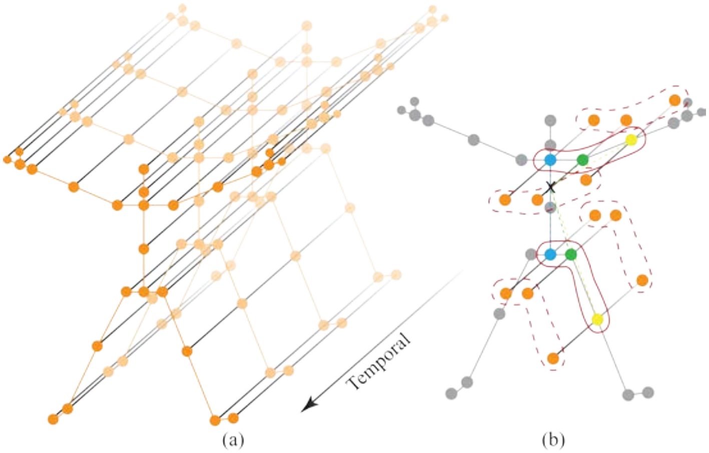


Figure 6.2 (a). Illustration of the spatiotemporal graph used in ST-GCN. (b). Illustration of the spatial mapping strategy. Different colors denote different subsets. × represents the center of gravity of the skeleton.

6.2.1.2 Graph Convolutional Operation

To represent the sampling area of convolutional operations, a neighbor set of a node υ_{ti} is defined as $\mathrm{N}(\upsilon_{ti}) = \{\upsilon_{tj} | d(\upsilon_{ti}, \upsilon_{tj}) \leq D\}$, where D is the minimum path length of $d(\upsilon_{ti}, \upsilon_{tj})$. Figure 6.2b shows this strategy, where × represents the center of gravity of the skeleton. The sampling area $\mathrm{N}(\upsilon_{ti})$ is enclosed by the curve. In detail, the strategy empirically uses 3 spatial subsets: the vertex itself (the green circle in Figure 6.2b); the centrifugal subset that contains the neighboring vertexes being farther from the gravity center (the yellow circle); the centripetal subset that contains the neighboring vertexes being closer to the center of gravity (the blue circle). Suppose there are fixed number of K subsets in the $\mathrm{N}(\upsilon_{ti})$, every neighbor set will be labelled numerically with a mapping $l_{ti}: \mathrm{N}(\upsilon_{ti}) \rightarrow \{0, \dots, K-1\}$. Temporally, the neighborhood concept is extended to temporally connected joints as $\mathrm{N}(\upsilon_{ti}) = \{\upsilon_{qj} | d(\upsilon_{tj}, \upsilon_{ti}) \leq K, |q-t| \leq \Gamma/2\}$, where Γ is the temporal kernel size that controls the temporal range of the neighbor set. Then the graph convolution could be computed as

$$f_{\text{out}}（\upsilon_{tj}） = \sum_{\boldsymbol{\upsilon_{tj}} \in \mathbf{N}(\upsilon_{ti})} \frac{1}{Z_{ti}(\upsilon_{tj})} f_{\text{in}}(\upsilon_{tj}) \mathrm{W}(l(\upsilon_{tj})) \tag{6.1}$$

where $f_{\text{in}}(\upsilon_{tj})$ is the feature map that get the attribute vector of υ_{tj}, $\mathrm{W}(l(\upsilon_{tj}))$ 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. $Z_{ti}(\upsilon_{tj}) = |\{\upsilon_{tk} | l_{ti}(\upsilon_{tk}) = l_{ti}(\upsilon_{tj})\}|$ is a normalization term that equals to the cardinality of the corresponding subset.

6.2.1.3 Implementation

The implementation of graph-based convolution is not as straightforward as 2D or 3D convolution. The feature map of the network could be represented by a tensor of (C, T, V) dimensions, where V denotes the number of vertexes, T denotes the temporal length and C denotes the number of attributes of the joint vertex. With the specific partitioning strategy determined, it could be represented by an adjacency matrix $\mathbf{A}$ with its elements indicating if a vertex $\boldsymbol{v}_{tj}$ belongs to a subset of $\mathbf{N}(v_{ti})$. The graph convolution is implemented by performing a $1 \times \Gamma$ classical 2D convolution and multiplies the resulting tensor with the normalized adjacency matrix $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{A}\mathbf{\Lambda}^{-\frac{1}{2}}$ on the second dimension. With K partitioning strategies $\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathbf{A}_{\mathrm{k}}$, Equation (6.1) could be transformed into

$$f_{\mathrm{out}}(\boldsymbol{v}_t) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathbf{\Lambda}_{\mathrm{k}}^{-\frac{1}{2}} \mathbf{A}_{\mathrm{k}} \mathbf{\Lambda}_{\mathrm{k}}^{-\frac{1}{2}} f_{\mathrm{in}} \mathrm{W}_{\mathrm{k}} \odot \mathrm{M}_{\mathrm{k}} \tag{6.2}$$

where $\mathbf{\Lambda}_k^{ii} = \sum_{\boldsymbol{j}} (\mathbf{A}_k^{ij}) + \alpha$ is a diagonal matrix with α set to 0.001 to avoid empty rows. W_{k} is a weight tensor of the 1×1 convolutional operation with $(C_{in}, C_{out}, 1, 1)$ dimensions, which represents the weighting function of Equation 6.1. M_{k} is an attention map with the same size of $\mathbf{A}_{\mathrm{k}}$, which indicates the importance of each vertex. $\odot$ denotes the element-wise product between two matrixes.

6.2.2 Architecture of 2T-GCN

The convolution for the temporal dimension follows the structure of ST-GCN, i.e., performing the $1 \times \Gamma$ convolution on the C×T×N feature maps. Both the spatial GCN and temporal GCN are followed by a batch normalization (BN) layer and a ReLU layer. As Figure 6.3 shows, one basic ST-GCN block is a stacked combination of one spatial GCN (Convs), one temporal GCN (Convt) and a dropout layer with a drop rate set of

0.5 to prevent overfitting. To stabilize the training progress, a residual connection is added at the end of each block.

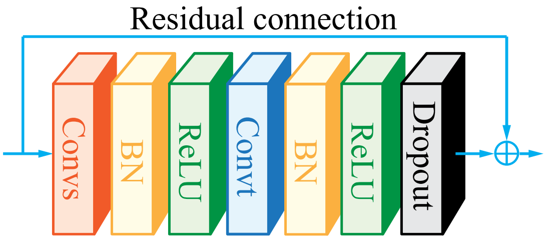


Figure 6.3 Structure of the ST-GCN block. Convs is the spatial GCN, and Convt is the temporal GCN, both of which are stacked with a BN layer and a ReLU layer. A residual connection is added at the end of block.

As Figure 6.3 shows, a basic GCN block is stacked with a spatial GCN (Convs) layer, a temporal GCN (Convt) layer and an additional dropout layer with the drop rate set to 0.5 to prevent overfitting. The convolution for the temporal dimension ensembles the implementation in [113], i.e., performing a $1 \times \Gamma$ convolution on the skeleton inputs. Both Convs and Convt are followed with a batch normalization (BN) layer and a ReLU layer. To stabilize the training, a residual connection is added for the GCN block. The proposed 2T-GCN model is constructed by a stack of these basic blocks. There are 9 such GCN blocks in total as shown in the middle part of Figure 7. The first three blocks, the middle three blocks, and the last three blocks have 64, 128 and 256 output channels, respectively. for. The temporal kernel size of these GCN blocks are set to 9. The strides of the 4-th and the 7-th GCN blocks are set to 2. To normalize the input data, a batch normalization layer is added at the beginning. As common practise in DL, we add a global average pooling layer at the end of the GCN blocks to transform the data to a vector with 256 dimensions. As the data are binarily labelled as normal or abnormal, we use a 2D convolutional layer to transform the 256-dimensional feature vector to a 2-dimensional vector. Finally, we feed the 2-dimensional vector to a SoftMax classifier to infer the abnormality of an exercise.

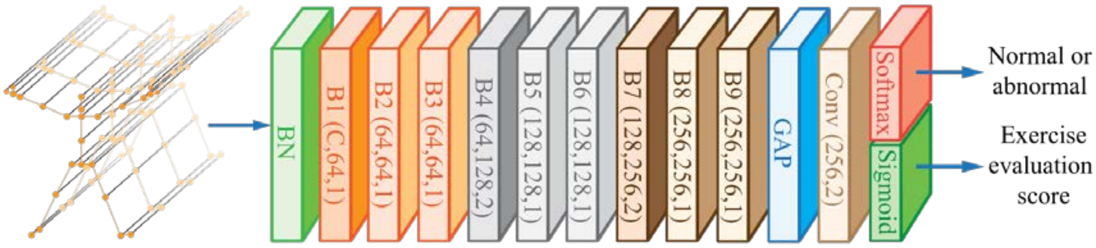


Figure 6.4 Illustration of the 2T-GCN model. There are 9 GCN blocks (B1-B9). The three numbers of each layer indicate the numbers of input and output channels, and the stride. GAP is a global average pooling layer.

To infer the exercise quality with a numerical evaluation score, we retrieve the probability distribution before the SoftMax layer of the model and transfer it to a range of [0,1] with a sigmoid function as

$$f_{\text{score}}(S^{(i)}) = \frac{1}{1 + e^{-f_{out}(S^{(i)})}} \tag{6.3}$$

The numerical score could indicate the exercise quality without the supervision of subjective human evaluation or arbitrary scores calculated by an function as in [153].

6.2.3 Optimization

To learn the weights Θ of the GCN model G, we defined the objective supervised by the binary clinical label y with the cross-entropy loss as

$$\arg\max_{\Theta} \sum_{i=1}^{N} - \sum y^{(i)} log(\sigma(G(\Theta, S^{(i)}))) \tag{6.4}$$

where $G(\Theta, S^{(i)})$ is the defined graph convolutional operation that is defined in Equation 6.2. σ represents the SoftMax function which transfers the probability distribution results to the abnormality result.

6.3 Results for HAE

In this section, we introduce the detailed implementation of our HAE algorithm on two datasets in terms of both abnormality detection and exercise quality evaluation.

6.3.1 Datasets

6.3.1.1 UI-PRMD Dataset

The UI-PRMD dataset [24] consists of skeletal data collected from 10 healthy subjects with every subject performing 10 repetitions of 10 rehabilitation exercises like deep squat, hurdle step, and sit to stand. The subjects performed every exercise both in a correct manner and in an incorrect manner, i.e., simulating performance by patients with musculoskeletal constraints. The data were acquired with two types of sensors namely Kinect v2 and Vicon optical tracking system that both provide position (3-D cartesian coordinates) and orientation (angular data) features of skeleton joints. As the dataset has inconsistent data caused by measurement errors and subjects performing the exercise with incorrect limbs, the dataset was then transferred to a consistent version by [75]. We use the consistent data from Kinect and Vicon sensors that both have 1326 repetitions. The purpose of using this dataset, on one hand, is to compare the HAE ability of our model with the one proposed by [75]. On the other hand, we investigate the effect of using different sensors to validate the properness of the Kinect v2 sensor used in this job.

6.3.1.2 Nursing Home Dataset

As described in the Section 3.4.3, our nursing home dataset has totally 869 repetitions that are performed by 25 elderly people. One task on the nursing home dataset is to further investigate which sensor features will mostly contribute to the abnormality detection. We use 8 features as described in Section 4.1 with their varied combinations to explore which of them will be the best for evaluating which exercise. Another goal is to examine whether our HAE method is consistent with the clinical evaluation.

6.3.2 Implementation Details

For training the proposed 2T-GCN mode, we adopt the same experimental setting for both UI-PRMD and EHE datasets. Precisely, we use the stochastic gradient descent to optimize our 2T-GCN model by setting the initial learning rate to 0.1. At the epochs of 10, 50 and 100, we decay the learning rate by multiplying it by 0.1. The training process will be terminated once the model achieves 100% accuracy, or it will stop at the epoch 200. The batch is set to 16. All experiments in are performed on a workstation with 2 GTX 1080 Ti GPUs.

6.3.3 Evaluation Criterion

To test the representation power of the model, we adopt the concept of separation degree (SD) that is proposed in [75]. For a pair of positive numbers x and y, their SD could be defined as $S_D(x, y) = \frac{x-y}{x+y} \in [-1,1]$. Then the separation degree between two positive sequences $\boldsymbol{x} = (x_1,\ x_2, \dots, x_m)$ and $\boldsymbol{y} = (y_1,\ y_2, \dots, y_m)$ could be defined by

$$S_D(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} S_D(x_i, y_j) \tag{6.5}$$

Meanwhile, we also use the distance metric defined in [154], which quantifies the difference between the correct and incorrect evaluation results. Given two positive sequences $\boldsymbol{x} = (x_1,\ x_2, \dots, x_N)$ and $\boldsymbol{y} = (y_1,\ y_2, \dots, y_N)$, the distance metric could be calculated as

$$D_n(x_n, y_n) = \frac{|x_n - y_n|}{\sqrt{\frac{1}{N} \sum_{n=1}^{N} (x_n - y_n)^2}} \tag{6.6}$$

For HAE, we also examine the activity evaluation ability by investigating whether the evaluation results are consistent with the clinic diagnostic results. To do so, we calculate the Euclidean Distance (ED) and correlation (CORREL) between the HAE results with the clinical labels. For an n-dimensional space, the distance of two vectors $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ is

$$E_D(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (x_i, y_i)^2} \tag{6.7}$$

The correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$Correl(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n} (x - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{6.8}$$

where $\bar{x}$ and $\bar{y}$ are the average value of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Smaller $E_D(\boldsymbol{x}, \boldsymbol{y})$ and larger $Correl(\boldsymbol{x}, \boldsymbol{y})$ indicates the evaluation result is more consistent with the observation of human expert and vice versa.

6.3.4 Results and Analysis

6.3.4.1 UI-PRMD Dataset

We calculate the SD of each exercise by using different features of two sensors as shown in Table 6.1. The average SD of [153] for the inter-subjects case is 0.515, while our method achieved a separation degree of 0.673 by using the same angular features of the Vicon optical tracking system. From the SD results, we could see that our method achieves a significant improvement over the best model named Log-likelihood GMM in [153]. By using the angular features of Kinect v2 sensor, our method achieves an even higher separation degree of 0.782. This indicates Kinect v2 could be a more capable sensor for HAE as it is better than the Vicon optical tracking system in terms of both its 3D position and angular features.

Table 6.1 Separation Degree of Every Exercise of UI-PRMD

Exercise ID	**Separation Degree (Std. Deviation)**			
	Kinect		**Vicon**	
	3D Position	**Angular**	**3D Position**	**Angular**
E1	0.795 (0.083)	0.806 (0.159)	0.405 (0.101)	**0.913 (0.086)**
E2	0.763 (0.102)	**0.96 (0.065)**	0.719 (0.127)	0.897 (0.022)
E3	0.504 (0.221)	**0.922 (0.05)**	0.536 (0.106)	0.678 (0.124)
E4	0.317 (0.198)	**0.581 (0.165)**	0.019 (0.023)	0.253 (0.258)
E5	0.268 (0.205)	**0.796 (0.074)**	0.31 (0.127)	0.753 (0.197)
E6	0.746 (0.177)	**0.649 (0.224)**	0.425 (0.211)	0.473 (0.112)
E7	0.847 (0.035)	**0.803 (0.117)**	0.727 (0.203)	0.821 (0.071)
E8	0.677 (0.189)	**0.729 (0.118)**	0.58 (0.299)	0.359 (0.358)
E9	0.783 (0.081)	**0.799 (0.108)**	0.404 (0.085)	0.665 (0.17)
E10	0.232 (0.03)	0.864 (0.034)	0.269 (0.104)	**0.928 (0.058)**
Average	0.594 (0.134)	**0.782 (0.117)**	0.426 (0.138)	0.673 (0.148)

To further validate the proposed method, we calculate the distance metric of all exercises. For each exercise in UI-PRMD, the mean and standard deviation of the distance metric are illustrated in Table 6.2. Exercises E1 and E7 (i.e., “deep squat” and “standing shoulder abduction”) are evaluated by using an autoencoder neural network proposed in [154]. With our method, the distance metric results of E1 and E7 (0.929 and 0.882, respectively) are higher than that of the autoencoder neural network (0.872

and 0.870, respectively). Meanwhile, our evaluation results are also more stable as the standard deviations of the distance metric of our method (0.146 and 0.158 for E1 and E7, respectively) is smaller than that of the autoencoder neural network (0.433 and 0.425 for E1 and E7, respectively). Besides, the evaluation results of distance metric also consistently indicate that the Kinect v2 sensor is more capable than the Vicon optical tracking system.

Table 6.2 Distance Metric of Every Exercise of UI-PRMD

Exercise ID	**Distance Metric (Std. Deviation)**			
	Kinect		**Vicon**	
	3D Position	**Angular**	**3D Position**	**Angular**
E1	0.758 (0.262)	0.899 (0.153)	0.373 (0.388)	**0.929 (0.146)**
E2	0.809 (0.246)	**0.964 (0.118)**	0.75 (0.321)	0.646 (0.388)
E3	0.66 (0.303)	**0.883 (0.277)**	0.461 (0.384)	0.812 (0.167)
E4	0.585 (0.274)	**0.743 (0.19)**	0.007 (0.016)	0.591 (0.249)
E5	0.546 (0.247)	0.840 (0.204)	0.37 (0.375)	**0.848 (0.192)**
E6	0.869 (0.163)	0.777 (0.265)	0.599 (0.293)	**0.9 (0.159)**
E7	0.755 (0.32)	0.726 (0.295)	0.7 (0.288)	**0.882 (0.158)**
E8	0.75 (0.331)	**0.71 (0.253)**	0.55 (0.445)	0.682 (0.333)
E9	0.671 (0.03)	**0.822 (0.258)**	0.317 (0.514)	0.748 (0.24)
E10	0.22 (0.369)	0.831 (0.216)	0.299 (0.277)	**0.912 (0.205)**
Average	0.669 (0.252)	**0.819 (0.220)**	0.434 (0.329)	0.802 (0.218)

Figures 6.5 and 6.6 show a visualized view of the exercise quality evaluation values of “deep squat” and “standing shoulder abduction” in UI-PRMD by using the 3D position features of Kinect v2. It is noticeable that the correct and incorrect repetitions are clearly classified by using the exercise quality evaluation score transferred from the probability distribution before the SoftMax layer with Equation 6.3. According to the results in [153], the correct and incorrect pairs could not be clearly separated by their DL method as most of the incorrect repetitions get evaluation scores around 0.9 (given that 1 is the fully correct score). With our method, we could see that the scores

of incorrect repetitions are below 0.5, whereas the correct repetitions have scores that are over 0.5.

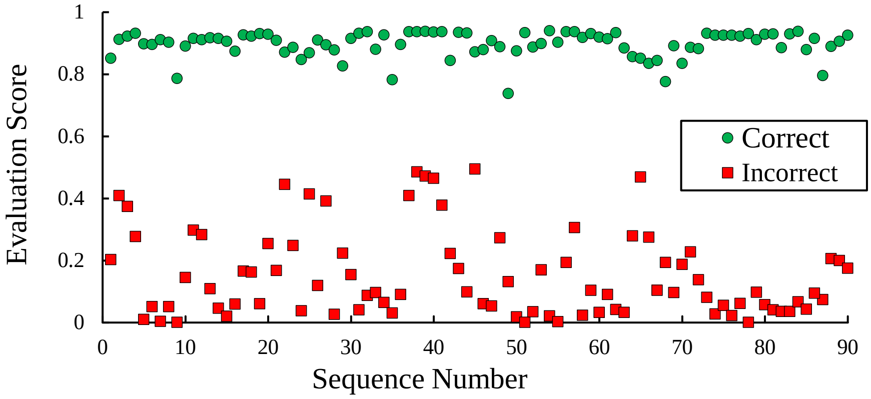


Figure 6.5 Quality evaluation scores of the “deep squat” ($S_D = 0.795$)

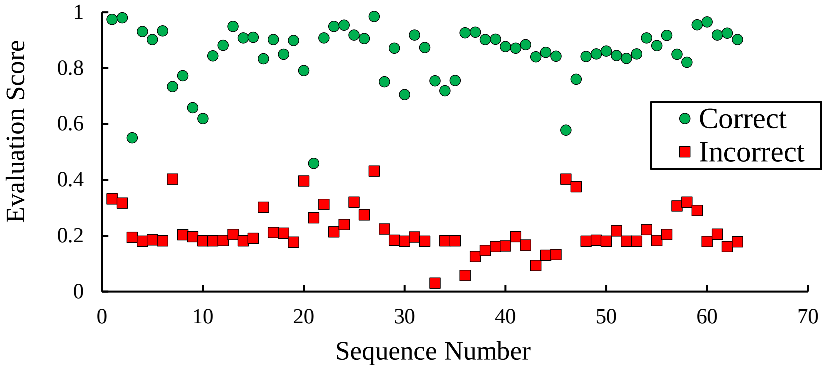


Figure 6.6 Quality evaluation scores of the “standing shoulder abduction” ($S_D = 0.847$)

6.3.4.2 Nursing Home Dataset

Since the Kinect v2 provides both the position and orientation attributes for the skeleton joints, we investigate the capability of different attributes with varied combinations of them. As Table 6.3 shows, experiments on 5 different features combinations are conducted by comparing their training accuracy on different exercises. In table 6.3, xyz refers to the 3D position attributes $(j_{ix}^t, j_{iy}^t, j_{iz}^t)$, xyzh refers to $(j_{ix}^t, j_{iy}^t, j_{iz}^t, j_{ih}^t)$, angular refers to $(j_{iX}^t, j_{iY}^t, j_{iZ}^t)$, angw refers to $(j_{iX}^t, j_{iY}^t, j_{iZ}^t, j_{iW}^t)$, and xyzhangw represents using all the attributes. The results indicate that some features could be good for the evaluation of some specific exercises. For example, xyz is good for the top 4 exercises but not good for exercises 5 (E5). The average accuracy of all attribute combinations is over 90%, which verifies the representation power of our HAE method. Since the model is a deep learning method, we feed all the attributes to the model to explore if it could automatically learn the discriminative feature from the data. It turns out the model have achieved promising results by using all the attributes of skeleton joints with 100% training accuracies on all exercises.

Table 6.3 Training Accuracy on UI-PRMD Dataset

Exercise ID	Training Accuracy (%)				
	angular	angw	xyz	xyzh	xyzhangw
E1	100.00	100.00	100.00	100.00	100.00
E2	99.31	100.00	100.00	100.00	100.00
E3	100.00	100.00	100.00	100.00	100.00
E4	100.00	100.00	100.00	95.70	100.00
E5	86.49	89.19	56.76	66.22	100.00
E6	100.00	100.00	98.59	85.92	100.00
Average	97.63	98.20	99.72	91.31	**100.00**

We also perform prediction experiments to evaluate if our method could be used to predict action abnormality that could reflect the severity of the AD. To make the experiments less biased than simply splitting the data to training set and test set, we adopt k-fold cross-validation evaluation method, which is popularly used to estimate the skill of a machine learning model when the sample size is relatively not large [119]. We set k to 5 for Cross-Validation (CV) by splitting the data of each exercise repetitions to 5 folds based on two types of evaluation criteria: cross-subjects and random-division. The of CV folds of two evaluation criteria are detailed in Table 6.4. For random-division evaluation, the remainder repetitions of the whole repetitions divided by 5 are also evenly put to each CV fold based on the fold order.

Table 6.4 Cross-validation folds of two evaluation types

CV Fold	Subject ID	
	Cross-Subjects	Random Division
Fold 1	2,3,1,4,7	Evenly divide the exercise repetitions of every subjects to 5 folds.
Fold 2	5,6,8,9,16	
Fold 3	10,11,17,18,23	
Fold 4	12,13,20,21,24	
Fold 5	14,15,19,22,25	

According to the comparison of different attribute combinations in Table 6.3, we use all attributes of the skeleton joints for the abnormality prediction. We report the abnormality prediction accuracy in percentage. Table 6.5 and Table 6.6 show the experimental results of the evaluation criteria of cross-subjects and random-division, respectively. The results consistently indicate that exercises like “wave hands” and “bend waist to right” could effectively reflect the Alzheimer severity, which is slightly better than the performance of exercises like “hands up and down” and “bend waist to right”. However, walking related exercises could not perform well for Alzheimer related abnormality prediction.

Table 6.5 Abnormality Prediction for Cross-Subjects Evaluation

CV Fold	Exercise Abnormality Prediction (%)					
	E1	E2	E3	E4	E5	E6
Fold 1	100.00	88.89	81.48	90.00	57.14	57.14
Fold 2	97.06	74.07	82.86	92.31	75.00	75.00
Fold 3	97.56	86.67	76.19	91.43	75.00	62.50
Fold 4	86.67	80.00	73.08	96.15	62.50	80.00
Fold 5	95.56	83.33	85.00	83.72	62.50	71.43
Avg	95.37	82.59	79.72	90.72	66.43	69.21

Table 6.6 Abnormality Prediction for Random-Division Evaluation

CV Fold	Exercise Abnormality Prediction (%)					
	E1	E2	E3	E4	E5	E6
Fold 1	98.08	82.50	82.00	91.49	60.00	60.00
Fold 2	97.83	89.66	88.10	95.12	62.50	70.83
Fold 3	95.00	88.00	87.18	97.30	63.16	72.22
Fold 4	90.62	84.00	88.24	96.97	-	-
Fold 5	96.43	84.00	87.10	96.43	-	-
Avg	95.59	85.63	86.52	95.46	61.89	67.68

To have a better intuition for the effectiveness of the proposed method, we visualize the average prediction scores of all exercises for 25 subjects in Figures 6.7 and 6.8 (different shaped markers) corresponding to the two evaluation protocols (i.e., cross-subjects division and random division, respectively). From the lines in Figures 6.7 and 6.8, we could observe that the prediction results of non-walking related exercises are consistent with clinical severity evaluations of AD under both two evaluation protocols. Our method also indicates there might be no discriminative features from walking related exercises like "walking forward" and "walking backward". This result could provide a guidance for designing exercises in the rehabilitation therapies as it is understandable that following basic actions like "walking forward" and "walking backward" rely less on the praxis condition. Given that our data is naturally collected instead of like UI-PRMD [24] which is collected with young subjects mimicking the abnormal exercises, our job reflects the real situation of Alzheimer Patients. On one hand, the results could imply that the Alzheimer patients usually have unnoticeable abnormal symptoms from the regular walking exercises. On the other hand, the proposed method is potential to capture severity levels of AD from exercises that require good praxis conditions like imitation, production, or recognition of gesture.

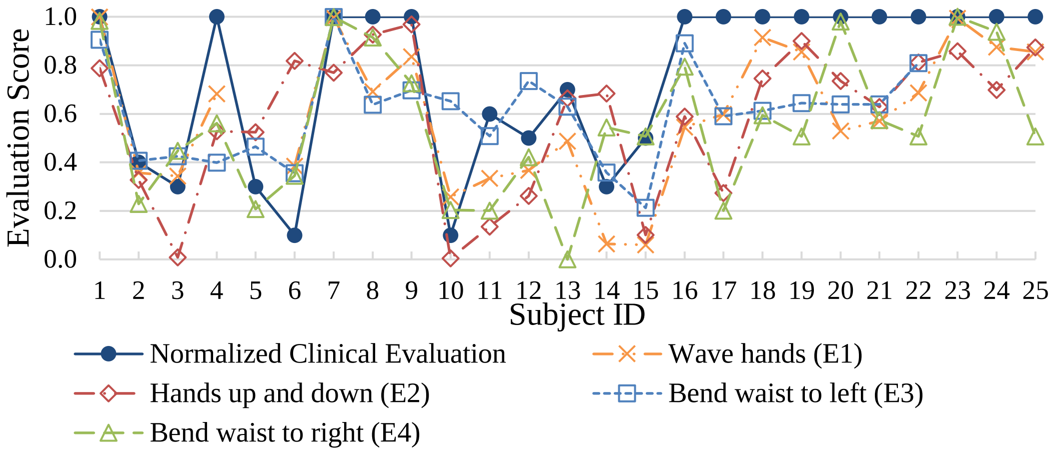


Figure 6.7 Average evaluation score of exercise 1 for all subjects by using different attribute combinations

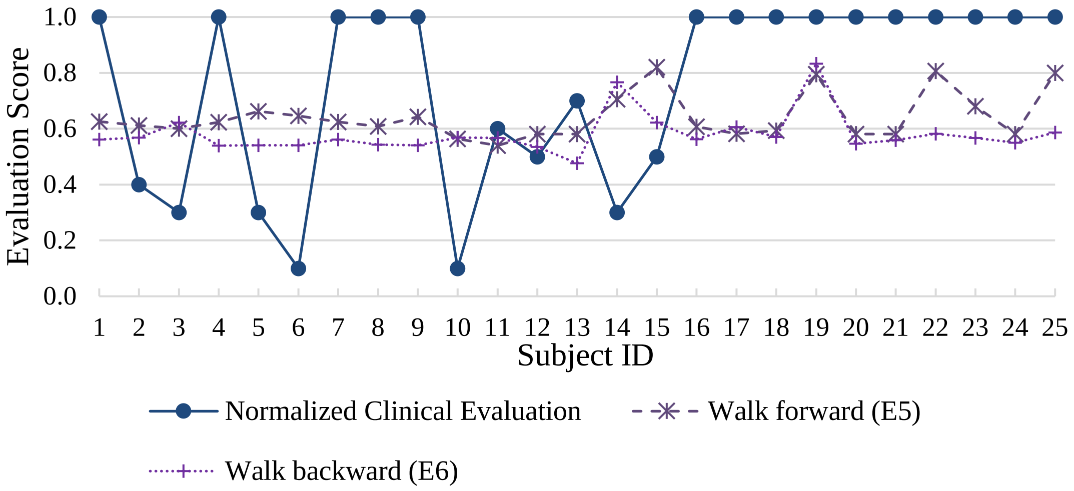


Figure 6.8 Average evaluation score of different exercises for all subjects by using all attributes of skeleton joints

To quantify the performance of our evaluation method, we use two association parameters defined as ED and CORREL in Equation 6.7 and Equation 6.8, respectively. We focus on investigating which exercises with which skeleton data attributes could achieve the best exercise evaluation performance by comparing their association with the clinical severity labels of AD. By considering both normal and Alzheimer subjects, we could see from Figure 6.9 that E1 (i.e., wave hands) could be the best exercise for inferring the abnormality as its evaluation scores has the lowest ED and highest correlation with clinical evaluation. In terms of evaluating the Alzheimer severity, according the correlation and normalized ED in Figure 6.10, it remains challenging to model the exercise evaluation score that is highly associated with the clinical observation. It is worthwhile to mention that the results of E1 could be at least consistent and positively associated with the clinical evaluation of the Alzheimer severity, which validates the motivation of this study.

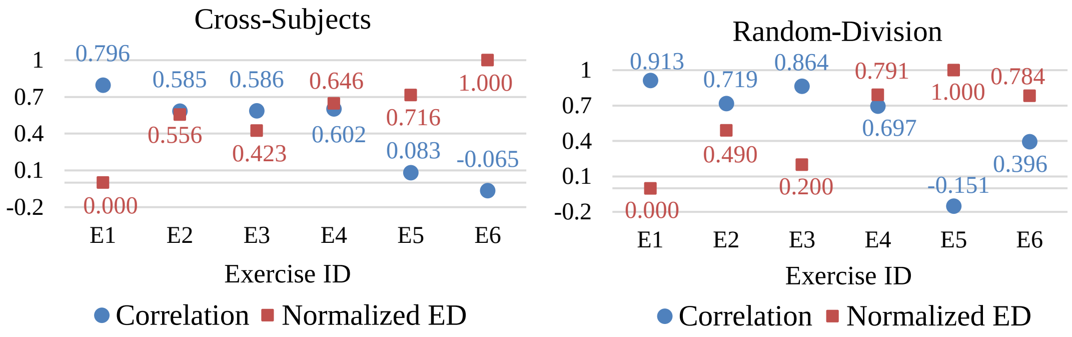


Figure 6.9 Comparison of HAE ability of different exercises considering all subjects (numbers are colored)

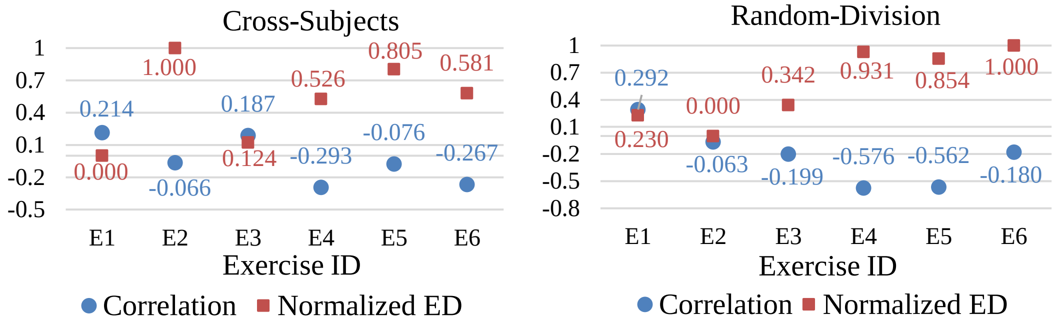


Figure 6.10 Comparison of HAE ability of different exercises considering Alzheimer subjects (numbers are colored)

6.4 Discussion

The proposed 2T-GCN solves the skeleton-based HAE from a novel perspective by modelling the problem as a binary classification problem. While existing methods usually model the skeleton-based HAE as a regression problem that are supervised either by arbitrarily predefined function scores or subjective human label. The validated 2T-GCN shows the potential to be applied to wide application domains like behavioral therapy and physical rehabilitations based on the condition that the skeleton data of actions could be well segmented. Hence, to apply it to real applications, the segmentation problem also needs to be tackled.

Chapter 7

Conclusion

7.1 Summary

This thesis introduced practical methods for human activity analysis tasks with competitive HAR and HAE algorithms proposed. The research is based on a comprehensive understanding of various sensors and domain requirements in the healthcare. Precisely, we investigated the related works of HAR in terms of sensors and datasets for the development of real-world ADLs recognition methods. An activity complexity definition method is provided and applied to evaluate the capability of various sensors for human activity analysis tasks. With such a comprehensive understanding, we utilized the Kinect v2 sensor and adopted various algorithms 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 with out proposed models. Our ABFE algorithm significantly improves the Top-1 accuracy comparing with other DL models. While our transfer learning method also improves the accuracy with enhanced efficiency and robustness. It is also worth mentioning that our MMNet achieved state-of-the-art performances on three public datasets. With such promising results, our proposed HAR algorithms are of great potential to be applied to real-world healthcare application scenarios like habit perception, intervention performance evaluation, disease prediction, and adaptive (automatic) smart home.

Except HAR, this thesis also provided a real-world field study that explores the potential of our proposed 2T-GCN for both diagnosis and therapy of Alzheimer's disease. For diagnosis, 2T-GCN could perform abnormality detection. Meanwhile, for exercise evaluation, the result of the abnormality detection could be transformed to a continuous score that indicates the elderly people's wellness of praxis condition. The proposed exercise evaluation model was first validated on the benchmark dataset named UI-PRMD, which significantly improves the results of [75] in terms of the separation degree and [154] in terms of distance metrics. Meanwhile, the results indicate that the Kinect v2 sensor is more capable for HAE tasks than the motion capture sensor called Vicon optical capture system. In addition to the laboratory-collected dataset, we also collected a real-world morning exercise dataset with real Alzheimer subjects in a nursing home. The experimental results on the real-world dataset that we collected in the nursing home show that our 2T-GCN is capable of discriminating abnormality in exercises, which could be used for supporting the diagnosis of AD. Meanwhile, the continuous evaluation score is also well associated with the Alzheimer severity of clinical observation, which indicates that our method could be used to monitoring the progress of exercise-based interventions.

7.2 Future Work

Although we achieved high accuracy on the morning routine dataset with three proposed algorithms, more research is required to investigate advanced HAR methods that provides more detailed information for the wide applications of HAR. 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 nursing homes and independent elderlies. Although medical datasets for various diseases are widely available, the lack of behavior datasets remains an issue for preventing elderly suffering from NCDs. As far as we know, there are very few field studies last for a long period and conduct a long-term HAR based disease evaluation. Given the high accuracy of our HAR methods on the ADLs dataset collected in the real world environment, one of our future work 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 of NCDs.

In terms of designing even more practical algorithms that make machine recognize actions like human, algorithms in the regimes of reinforcement learning [155], lifelong learning [156], self-training or active learning [157] could be utilized to handle problems like catastrophic forgetting, incremental actions, changing transition dynamics, changing rewards functions, etc. Meanwhile, new evaluation criteria with consideration of the improved QoL perception of users should also be proposed to target higher goals of HAR applications. For example, feedback from users could be surveyed when the proposed algorithms are used in the wide applications.

On the other hand, our 2T-GCN model shows the potential for detecting the severity of the Alzheimer’s disease with a numerical evaluation score, it might be lack of the involvement of domain knowledge as it could not be well explained. In the future, we will focus on this issue to build an explainable model and expand the experiments to a larger dataset. To do so, we will develop future HAE methods with inter-rater validation by comparing the evaluation results with other clinical scales and biomarkers.

References

[1] Q. Lei, J.-X. Du, H.-B. Zhang, S. Ye, and D.-S. Chen, "A Survey of Vision-Based Human Action Evaluation Methods," *Sensors,* vol. 19, no. 19, p. 4129, 2019.

[2] U. DESA, "World Population Prospects 2019: Highlights," *New York (US): United Nations Department for Economic and Social Affairs,* 2019.

[3] J. Gill and M. J. Moore, "The State of aging & health in America 2013," 2013. [Online]. Available: https://www.statista.com/statistics/207347/causes-of-death-among-us-adults-aged-65-by-ethnicity/.

[4] S.-R. Ke, H. 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.

[5] A. L. Martin-Niedecken and U. Götz, "Design and evaluation of a dynamically adaptive fitness game environment for children and young adolescents," in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts,* 2016: ACM, pp. 205-212.

[6] J. C. Chan, H. Leung, J. K. Tang, and T. Komura, "A virtual reality dance training system using motion capture technology," *IEEE Transactions on Learning Technologies,* vol. 4, no. 2, pp. 187-195, 2011.

[7] J. E. Deutsch *et al.*, "Nintendo wii sports and wii fit game analysis, validation, and application to stroke rehabilitation," *Topics in stroke rehabilitation,* vol. 18, no. 6, pp. 701-719, 2011.

[8] S. Ranasinghe, F. Al Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," *International Journal of Distributed Sensor Networks,* vol. 12, no. 8, p. 1550147716665520, 2016.

[9] T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz, and G. Tröster, "Wearable activity tracking in car manufacturing," *IEEE Pervasive Computing,* vol. 7, no. 2, 2008.

[10] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer,* vol. 29, no. 10, pp. 983-1009, 2013.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[18] 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.

[19] 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.

[20] 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.

[21] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3D human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010-1019.

[22] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 2019.

[23] D. González-Ortega, F. Díaz-Pernas, M. Martínez-Zarzuela, and M. Antón-Rodríguez, "A Kinect-based system for cognitive rehabilitation exercises monitoring," *Computer methods and programs in biomedicine*, vol. 113, no. 2, pp. 620-631, 2014.

[24] A. Vakanski, H.-p. Jun, D. Paul, and R. Baker, "A data set of human body movements for physical rehabilitation exercises," *Data*, vol. 3, no. 1, p. 2, 2018.

[25] 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.

[26] 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.

[27] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A Public Domain Dataset for Human Activity Recognition using Smartphones," in *ESANN*, 2013.

[28] 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.

[29] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2649-2656.

[30] S. Gasparrini, E. Cippitelli, S. Spinsante, and E. Gambi, "A depth-based fall detection system using a Kinect® sensor," *Sensors*, vol. 14, no. 2, pp. 2756-2775, 2014.

[31] V. Elangovan, V. K. Bandaru, and A. Shirkhodaie, "Team activity analysis and recognition based on Kinect depth map and optical imagery techniques."

[32] 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.

[33] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for event and object recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3272-3279.

[34] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition," in *European Conference on Computer Vision*, 2014: Springer, pp. 742-757.

[35] Y. Ji, F. Xu, Y. Yang, F. Shen, H. T. Shen, and W.-S. Zheng, "A Large-scale RGB-D Database for Arbitrary-view Human Action Recognition," in *2018 ACM Multimedia Conference on Multimedia Conference*, 2018: ACM, pp. 1510-1518.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346-362, 2017.

[42] 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," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2117-2126.

[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] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *32nd AAAI conference on artificial intelligence*, 2018.

[45] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3595-3603.

[46] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 786-792.

[47] W. Zhu *et al.*, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *30th AAAI Conference on Artificial Intelligence*, 2016.

[48] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1227-1236.

[49] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-Based Action Recognition With Directed Graph Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7912-7921.

[50] A. Haque *et al.*, "Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance," *arXiv preprint arXiv:1708.00163*, 2017.

[51] W. Kay *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[52] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *2011 International Conference on Computer Vision*, 2011: IEEE, pp. 2556-2563.

[53] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725-1732.

[54] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961-970.

[55] H. Idrees *et al.*, "The THUMOS challenge on action recognition for videos “in the wild”," *Computer Vision and Image Understanding*, vol. 155, pp. 1-23, 2017.

[56] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299-6308.

[57] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 305-321.

[58] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4768-4777.

[59] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*, 2016: Springer, pp. 510-526.

[60] G. A. Sigurdsson, O. Russakovsky, and A. Gupta, "What actions are needed for understanding human actions in videos?," in *Computer Vision (ICCV), 2017 IEEE International Conference on,* 2017: IEEE, pp. 2156-2165.

[61] 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.

[62] 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.

[63] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in rgb+ d videos," *IEEE transactions on pattern analysis and machine intelligence,* 2017.

[64] D. Wu *et al.*, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 38, no. 8, pp. 1583-1597, 2016.

[65] 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.

[66] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Image Processing (ICIP), 2015 IEEE International Conference on,* 2015: IEEE, pp. 168-172.

[67] B. Pan, J. Sun, W. Lin, L. Wang, and W. Lin, "Cross-Stream Selective Networks for Action Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,* 2019, pp. 0-0.

[68] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2018, pp. 469-478.

[69] F. Baradel, C. Wolf, and J. Mille, "Human action recognition: Pose-based attention draws focus to hands," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 604-613.

[70] F. Baradel, C. Wolf, and J. Mille, "Human activity recognition with pose-driven attention to rgb," in *BMVC 2018 - 29th British Machine Vision Conference*, Newcastle, United Kingdom, 2018, pp. pp.1-14.

[71] M. A. R. Ahad, A. D. Antar, and O. Shahid, "Vision-based Action Understanding for Assistive Healthcare: A Short Review," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2019,* 2019, pp. 1-11.

[72] B. Galna, G. Barry, D. Jackson, D. Mhiripiri, P. Olivier, and L. Rochester, "Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease," *Gait & posture,* vol. 39, no. 4, pp. 1062-1068, 2014.

[73] Z. S. de Urturi Breton, B. G. Zapirain, and A. M. Zorrilla, "Kimentia: Kinect based tool to help cognitive stimulation for individuals with dementia," in *2012*

IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom), 2012: IEEE, pp. 325-328.

[74] L. Tao *et al.*, "A comparative study of pose representation and dynamics modelling for online motion quality assessment," *Computer vision and image understanding,* vol. 148, pp. 136-152, 2016.

[75] Y. Liao, A. Vakanski, and M. Xian, "A Deep Learning Framework for Assessing Physical Rehabilitation Exercises," *arXiv preprint arXiv:1901.10435,* 2019.

[76] F. Sardari, A. Paiement, and M. Mirmehdi, "View-Invariant Pose Analysis for Human Movement Assessment from RGB Data," in *International Conference on Image Analysis and Processing,* 2019: Springer, pp. 237-248.

[77] J. Antunes, A. Bernardino, A. Smailagic, and D. P. Siewiorek, "AHA-3D: A Labelled Dataset for Senior Fitness Exercise Recognition and Segmentation from 3D Skeletal Data," in *BMVC,* 2018, p. 332.

[78] E. Dove and A. Astell, "The Kinect Project: Group motion-based gaming for people living with dementia," *Dementia,* vol. 18, no. 6, pp. 2189-2205, 2019.

[79] X. Bruce and K. C. Chan, "Discovering Knowledge by Behavioral Analytics for Elderly Care," in *Big Knowledge (ICBK), 2017 IEEE International Conference on*, 2017: IEEE, pp. 284-289.

[80] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer vision and image understanding,* vol. 73, no. 3, pp. 428-440, 1999.

[81] 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.

[82] 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.

[83] 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.

[84] A. Wickramasinghe, D. C. Ranasinghe, C. Fumeaux, K. D. Hill, and R. Visvanathan, "Sequence learning with passive RFID sensors for real-time bed-egress recognition in older people," *IEEE journal of biomedical and health informatics,* vol. 21, no. 4, pp. 917-929, 2017.

[85] R. L. S. Torres, D. C. Ranasinghe, Q. Shi, and A. P. Sample, "Sensor enabled wearable RFID technology for mitigating the risk of falls near beds," in *RFID (RFID), 2013 IEEE International Conference on,* 2013: IEEE, pp. 191-198.

[86] 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.

[87] 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.

[88] 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.

[89] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *arXiv preprint arXiv:1611.08050,* 2016.

[90] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017: IEEE, pp. 1561-1570.

[91] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017: IEEE, pp. 1263-1272.

[92] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014: Springer, pp. 740-755.

[93] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Computer Vision and Image Understanding*, vol. 158, pp. 85-105, 2017.

[94] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," in *Image Processing (ICIP), 2014 IEEE International Conference on*, 2014: IEEE, pp. 1565-1569.

[95] "Kinect for Windows SDK v1.8." https://www.microsoft.com/en-us/download/details.aspx?id=40278 (accessed.

[96] "OPENNI SDK HISTORY." http://openni.ru/openni-sdk/openni-sdk-history-2/index.html (accessed.

[97] "Kinect tools and resources." https://developer.microsoft.com/en-us/windows/kinect/tools (accessed.

[98] M. WILSON. "Exclusive: Microsoft Has Stopped Manufacturing The Kinect." https://www.fastcodesign.com/90147868/exclusive-microsoft-has-stopped-manufacturing-the-kinect (accessed.

[99] "The OpenNI Library." https://structure.io/openni (accessed.

[100] "Leap Motion SDK v2." https://developer.leapmotion.com/get-started/ (accessed.

[101] "Intel® RealSense™ SDK for Windows." https://software.intel.com/en-us/realsense-sdk-windows-eol (accessed.

[102] "Intel® RealSense™ SDK 2.0." https://software.intel.com/en-us/realsense/sdk (accessed.

[103] 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.

[104] 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.

[105] J. Xu, K. D. Kochanek, S. L. Murphy, and B. Tejada-Vera, "Deaths: final data for 2014," 2016.

[106] "Seven Stages of Dementia | Symptoms & Progression." Dementia Care Central. https://www.dementiacarecentral.com/aboutdementia/facts/stages/ (accessed.

[107] M. S. Albert *et al.*, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging ‐ Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 270-279, 2011.

[108] J. S. Lin, E. O'Connor, R. C. Rossom, L. A. Perdue, and E. Eckstrom, "Screening for cognitive impairment in older adults: a systematic review for

the US Preventive Services Task Force," *Annals of internal medicine,* vol. 159, no. 9, pp. 601-612, 2013.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *arXiv preprint arXiv:1801.07455,* 2018.

[114] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data,* vol. 3, no. 1, p. 9, 2016.

[115] R. Rojas, "AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting," *Freie University, Berlin, Tech. Rep,* 2009.

[116] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate lstm-fcns for time series classification," *arXiv preprint arXiv:1801.04503,* 2018.

[117] A. Graves, "Supervised sequence labelling," in *Supervised sequence labelling with recurrent neural networks*: Springer, 2012, pp. 5-13.

[118] 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.

[119] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013.

[120] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote sensing of Environment,* vol. 62, no. 1, pp. 77-89, 1997.

[121] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[122] 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.

[123] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291-7299.

[124] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12026-12035.

[125] L. Zhang, G. Zhu, L. Mei, P. Shen, S. A. A. Shah, and M. Bennamoun, "Attention in convolutional LSTM for gesture recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 1953-1962.

[126] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation

nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802-810.

[127] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588-595.

[128] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5344-5352.

[129] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2017, pp. 1647-1656.

[130] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5323-5332.

[131] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489-4497.

[132] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *European Conference on Computer Vision*, 2016: Springer, pp. 203-220.

[133] B. Li, H. Chen, Y. Chen, Y. Dai, and M. He, "Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2017: IEEE, pp. 613-616.

[134] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW),* 2017: IEEE, pp. 597-600.

[135] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based lstm networks for 3d action recognition and detection," *IEEE Transactions on Image Processing,* vol. 27, no. 7, pp. 3459-3471, 2018.

[136] W. Nie, W. Wang, and X. Huang, "SRNet: Structured Relevance Feature Learning Network From Skeleton Data for Human Action Recognition," *IEEE Access,* vol. 7, pp. 132161-132172, 2019.

[137] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2015, pp. 1110-1118.

[138] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1012-1020.

[139] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: IEEE, pp. 1362-1369.

[140] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham, "3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2601-2608.

[141] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2458-2466.

[142] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," presented at the The 28th International Conference on Machine Learning, 2011.

[143] B. Reisberg, "Global measures: utility in defining and measuring treatment response in dementia," *International Psychogeriatrics,* vol. 19, no. 3, pp. 421-456, 2007.

[144] B. Reisberg, S. H. Ferris, A. Kluger, E. Franssen, J. Wegiel, and M. J. De Leon, "Mild cognitive impairment (MCI): a historical perspective," *International Psychogeriatrics,* vol. 20, no. 1, pp. 18-31, 2008.

[145] B. Dubois *et al.*, "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria," *The Lancet Neurology,* vol. 6, no. 8, pp. 734-746, 2007.

[146] A. Diamond, "Executive functions," *Annual review of psychology,* vol. 64, pp. 135-168, 2013.

[147] S. M. Arkin and N. Morrow-Howell, "Elder rehab: A student-supervised exercise program for Alzheimer's patients," *The Gerontologist,* vol. 39, no. 6, pp. 729-735, 1999.

[148] L. Teri *et al.*, "Exercise plus behavioral management in patients with Alzheimer disease: a randomized controlled trial," *Jama,* vol. 290, no. 15, pp. 2015-2022, 2003.

[149] C. W. Cotman, N. C. Berchtold, and L.-A. Christie, "Exercise builds brain health: key roles of growth factor cascades and inflammation," *Trends in neurosciences,* vol. 30, no. 9, pp. 464-472, 2007.

[150] P. J. Smith *et al.*, "Aerobic exercise and neurocognitive performance: a meta-analytic review of randomized controlled trials," *Psychosomatic medicine,* vol. 72, no. 3, p. 239, 2010.

[151] P. Heyn, B. C. Abreu, and K. J. Ottenbacher, "The effects of exercise training on elderly persons with cognitive impairment and dementia: a meta-analysis," *Archives of physical medicine and rehabilitation,* vol. 85, no. 10, pp. 1694-1704, 2004.

[152] A.-B. Haidich, "Meta-analysis in medical research," *Hippokratia,* vol. 14, no. Suppl 1, p. 29, 2010.

[153] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 28, no. 2, pp. 468-477, 2020.

[154] C. Williams, A. Vakanski, S. Lee, and D. Paul, "Assessment of physical rehabilitation movements through dimensionality reduction and statistical modeling," *Medical engineering & physics,* vol. 74, pp. 13-22, 2019.

[155] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[156] Y. Chandak, G. Theocharous, C. Nota, and P. S. Thomas, "Lifelong Learning with a Changing Action Set," in *AAAI*, 2020, pp. 3373-3380.

[157] C. Bettini, G. Civitarese, and R. Presotto, "CAVIAR: Context-driven Active and Incremental Activity Recognition," *Knowledge-Based Systems,* p. 105816, 2020.