Année 2017 Num d'ordre: 42267

Université Lille 1

Ecole Doctorale Science Pour l'Ingénieur

Laboratoire CRIStAL (UMR CNRS 9189)

Équipe 3D SAM

pour obtenir le grade de
Docteur,
spécialité Informatique

présentée et soutenue publiquement par

Meng Meng

le 9/01/2017

Human Object Interaction Recognition

Directeur de thèse
Mohamed DAOUDI
Encadrants
Hassen DRIRA
Jacques BOONAERT

COMPOSITION DU JURY

M. Stefano Berretti	Professeur, Università di Firenze	Rapporteur
M. Djamel Merad	Maître de conférences HDR, Université d'Aix-Marseille	Rapporteur
M. Mohamed Daoudi	Professeur, Télécom Lille	Directeur
M. Hassen Drira	Maître de conférences, Télécom Lille	Encadrant
M. Jacques Boonaert	Maître de conférences, École des Mines de Douai	Encadrant
M. Frédéric Lerasle	Professeur, Université Paul Sabatier	Examinateur
Mme. Catherine Soladie	Maître de conférences, CentraleSupélec	Examinatrice
M. Hichem Snoussi	Professeur, Université de technologie de Troyes	Examinateur
M. Christian Wolf	Maître de conférence HDR, INSA Lyon, laboratoire LIRIS	Invité

Abstract

In this thesis, we have investigated the human object interaction recognition by using the skeleton data and local depth information provided by RGB-D sensors. There are two main applications we address in this thesis: human object interaction recognition and abnormal activity recognition.

First, we propose a spatio-temporal modeling of human-object interaction videos for on-line and off-line recognition. In the spatial modeling of human object interactions, we propose low-level feature and object related distance feature which adopted on on-line human object interaction recognition and abnormal gait detection. Then, we propose object feature, a rough description of the object shape and size as new features to model human-object interactions. This object feature is fused with the low-level feature for online human object interaction recognition. In the temporal modeling of human object interactions, we proposed a shape analysis framework based on low-level feature and object related distance feature for full sequence-based off-line recognition. Experiments carried out on two representative benchmarks demonstrate the proposed method are effective and discriminative for human object interaction analysis.

Second, we extend the study to abnormal gait detection by using the on-line framework of human object interaction classification. The experiments conducted following state-of-the-art settings on the benchmark shows the effectiveness of proposed method.

Finally, we collected a multi-view human object interaction dataset involving abnormal and normal human behaviors by RGB-D sensors. We test our model on the new dataset and evaluate the potential of the proposed approach.

Key words: skeleton data, human object interaction recognition, spatio-temporal modeling, on-line recognition, abnormal gait, multi-view dataset, rate invariant recognition, trajectories analysis.

Résumé

Dans cette thèse, nous avons étudié la reconnaissance des actions qui incluent l'intéraction avec l'objet à partir des données du skeleton et des informations de profondeur fournies par les capteurs RGB-D. Il existe deux principales applications que nous abordons dans cette thÃ¨se: la reconnaissance de l'interaction humaine avec l'objet et la reconnaissance d'une activité anormale.

Nous proposons, dan un premier temps, une modélisation spatio-temporelle pour la reconnaissance en ligne et hors ligne des intéractions entre l'humain et l'objet. Dans la modélisation spatiale, nous proposons des caractéristiques de bas niveau liés à la distance entre les points du skeleton et la distance entre l'objet et les points du skeleton. Ces caractéristiques ont été adoptées pour la reconnaissance en ligne des intéractions humaines avec l'objet et pour la détection de la démarche anormale. Ensuite, nous proposons des caractéristiques liées à d'objet qui décrivent approximativement la forme et la taille de l'objet. Ces caractéristiques sont fusionnées avec les caractéristiques bas-niveau pour la reconnaissance en ligne des intéractions humaines avec l'objet. Dans la modélisation temporelle, nous avons proposé un framework élastique pour aligner les trajectoires des distances dans le temps afin de permettre une reconnaissance hors ligne invariante au taux d'exécution. Les expériences menées sur deux benchmarks démontrent l'efficacité de la méthode proposée. Dans le deuxième volet de ce travail, nous étendons l'étude à la détection de la démarche anormale en utilisant le cadre en ligne l'approche. Afin de valider la robustesse de l'approche à la pose, nous avons collecté une base multi-vues pour des intéractions humaines avec l'objet, de façon normale et anormale. Les résultats expérimentaux sur le benchmark des actions anormales frontales et sur la nouvelles base prouvent l'efficacité de l'approche.

Mots-clés: Intéraction humaine avec l'objet, base multi-vues, données skeleton, démarche anormale, modélisations spatio-temporelle, invariance au taux d'exécution, reconnaissance en ligne, analyse de trajectoires.

Acknowledgement

This research work has been realized at "Human object interaction recognition (CRIStAL)" in Université Lille1, with the research group 3D cam from September 2013 to December 2016. First and foremost I offer my sincerest gratitude to my PhD director, Prof. Mohamed Daoudi, for his supervision, valuable guidance, continuous encouragement as well as given me extraordinary experiences through out my Ph.D. experience. I could not have imagined having a better tutor and mentor for my Ph.D. study.

I also thank my co-supervisors Dr. Hassen Drira and Dr. Jacques Booneart for their time and advice in beneficial scientific discussions. Their friendly and encouraging guidance make me more confident in my research.

Special thanks to my PhD committee members for taking the time to participate in this process, and especially the reviewers of the manuscript for having accepted this significant task: Prof. Stefano Berretti and Prof. Djamel Merad. All these people made me the honor to be present for this special day despite their highly charged agendas.

I would like to take the opportunity to express my gratitude and to thank my fellow workmates in CRIStAL. All of them have given me support and encouragement in my thesis work.

A special acknowledgment should be shown to Prof. Yiding Wang at the North China University of Technology, who enlightened me at the first glance of research. I always benefit from the abilities that I obtained on his team.

Last but not least, I convey special acknowledgement to my parents, Lingqi Meng and Xiangling XU, for supporting me to pursue this degree and to accept my absence for four years of living abroad.

Lille, December 1, 2016.

Publications

International journals

- **Meng Meng**, Hassen Drira and Jacques Boonaert. Distances evolution analysis for online and off-line human object interaction recognition. *Image and Vision Computing* (Under review).

- **Meng Meng**, Hassen Drira and Jacques Boonaert. Human abnormal and normal gait detection.*Pattern Recognition Letters* (Submitted).

International conferences

- **Meng Meng**, Hassen Drira, Mohamed Daoudi and Jacques Boonaert. Human-Object Interaction Recognition by Learning the Distances between the Object and the Skeleton Joints. *IEEE International Conference on Automatic Face and Gesture Recognition (FG) Workshop*, Ljubljana, Slovenia, 2015.

- **Meng Meng**, Hassen Drira, Mohamed Daoudi and Jacques Boonaert. Detection of Abnormal Gait from Skeleton Data. *International Conference on Computer Vision Theory and Applications (VISAPP)*, Rome, Italy, 2016

- **Meng Meng**, Hassen Drira, Mohamed Daoudi and Jacques Boonaert. Human Object Interaction Recognition using Rate-Invariant Shape Analysis of Inter Joint Distances Trajectories. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Differential Geometry in Computer Vision and Machine Learning*, Las Vegas, USA, 2016

Contents

List of Figures

1 Introduction

Sommaire

Spatio-temporal human representation based on 3D visual perception data is a rapidly growing research area. Indeed, this represents a task of interest for a wide spectrum of areas due to its huge potential, like human-machine interaction, physical rehabilitation, surveillance security, health care and social assistance, video games.

Comparing to verbal or vocal communication data, visual data forms one of the most important cues in developing systems for understanding human behavior. The applications range are from tracking daily activities to classifying emotional states, as well as detecting abnormal and suspicious activities.

In addition to pose variation and scale variability, high complexity of human motions and the variability of object interactions represent additional significant challenges. However, human activity understanding is a more challenging problem due to the diversity and complexity of human behaviors and accurate human action recognition is still a quite challenging task and is gradually moving towards more structured interpretation of complex human activities involving multiple people and especially interaction with objects. Motivated by this issue, and the need for efficient algorithms we focus our study on dynamic human object interaction recognition in this thesis.

Imaging technologies have recently shown a rapid advancement with the introduction of low cost depth cameras with real-time capabilities, like Microsoft Kinect that changed the picture by providing 3D depth data of video-based human action recognition. Compared to standard cameras, these range sensors provide 3D structural information of the scene, which offers more discerning information to recover human postures. The infrared technology behind these sensors allows them to work in complete darkness and to be robust to light and illumination variation, a common issue in 2D videos analysis. The real time acquisition and the advantages of these data encourage its use in several applications that need to understand and to recognize human object interactions using such a data stream instead of 2D videos.

Nowadays a lot of literature put effort on the real-time estimation of human body joints from depth images. This kind of skeleton data includes a set of 3D connected joints representing various human body parts to facilitate the analysis of the human pose and its motion over the time. Many work already proved the effectiveness of skeleton data for the analysis and recognition of relatively simple behaviors like human gestures or actions.

Even though the depth cameras generally have better quality 3D action data than those estimated from monocular video sensors, adopting the 3D joint positions for human-object interaction is not sufficient to classify action especially including interaction with objects. Actually, during a human object interaction scene, the hands may hold objects and are hardly detected or recognized due to heavy occlusions and appearance variations. A high level of information of the objects is needed to recognize the human-object interaction. On the other hand, the use of 3D skeleton joints is not sufficient to distinguish some actions like *drinking* and *picking phone*. Extra inputs need to be included and exploited for more accurate recognition.

There are several advantages of this data: easy to remove background; isolating and tracking human body; allowing capturing the human motion at each frame. Additionally, the 3D depth data are independent of human appearance (textures) and provide a more complete human silhouette relative to the silhouette information used in the past. So the emergence of 3D depth data reduces the challenges to human behavior analysis.

Along With the release of the Kinect, effective methods that take advantage of body-joints information and depth video have been proposed. There are still many challenges in human object interaction recognition like rough location and shape of object, invariant to geometric transformations of the subject and different execution speed of the same action. All of these problems should be considered in an effective and robust human object interaction recognition system. Nevertheless, the real-time nature of the device bring another challenge: online recognition system especially in the human object interaction context is needed.

1.1 Motivation and Challenges

Human activity understanding is a more challenging problem due to the diversity and complexity of human behaviors [138] and accurate human action recognition is still a quite challenging task and is gradually moving towards more structured interpretation of complex human activities involving multiple people and especially interaction with objects. To the best of our knowledge, the majority of action recognition past approaches investigate simple action recognition [3] [108] [130] [99] [7] [128] such as boxing, kicking, walking, etc. and less effort have been spent on human object interaction. There are two challenges for human-object interaction recognition. The first one is online classification that need low level features, and the other scenario is to classify full videos. This scenario introduces a new challenge which is the difference in rate and execution time.

Even though the depth cameras generally have better quality 3D action data than those estimated from monocular video sensors, adopting the 3D joint positions for human-object interaction is not sufficient to classify action especially including interaction with objects. Actually, during a human object interaction scene, the hands may hold objects and are hardly detected or recognized due to heavy occlusions and appearance variations [119]. A high level of information of the objects is needed to recognize the human-object interaction. On the other hand, the use of 3D skeleton joints is not sufficient to distinguish some actions like *drinking* and *picking phone*. Extra inputs need to be included and exploited for more accurate recognition.

1.2 Thesis contributions

Motivated by all considerations stated above, this PhD thesis investigate the issue of human activity recognition using low-cost 3D senors with a focus on human-object interactions in three main applications; the abnormal gait detection and human-object interaction recognition within two different scenarios: online recognition and full sequence-based recognition. The main contributions summarized as follows:

- *human object interaction recognition:* We propose a framework to recognize human object interaction for two different scenarios. For online human-object interaction recognition, we propose to use original translation and rotation invariant representation of a new object feature describing roughly the size and the shape of the object to recognize human-object interaction. These features are used together with object-related features use low-level features as input to classifier. The proposed the distances between the objects and the human joints represent a rough description of the object shape and size as new features to model the human-object interactions. These features are fused with the low-level features (inter-joints) for human-object interaction recognition. This low-level feature are calculated in each frame and the sequence is modeled as evolution of the resulting feature vector; this step is denoted spatio-temporal modeling. The rough shape and the size of the object detection represent the next step in the pipeline of online human-object interaction recognition. Together with the low-level feature resulting on Spatio-temporal modeling, the object feature represent the input of the random forest classifier.

  Several applications require human object interaction recognition after the action is done. This scenario seems less constraining on real-time and rapidity of calculus. Furthermore, the full sequence-based human object interaction recognition seems easier scenario compared to online recognition. However, this scenario reveals additional challenges arise such as execution time differing for same interaction and significant spatial variation in the way of performing an action. A more elaborated spatio-temporal modeling is proposed here. The evolution of the inter-joints and object-joints distances in time is modeled as trajectories in a high dimension space and a shape analysis framework is used to analyze and compare the corresponding trajectories in a Riemannian manifold. This framework has the advantage to make the re-parameterization group acting by isometry on the space of these trajectories. The distance between the orbits corresponding to two trajectories is invariant to the rate of execution in the sequence. Another advantage of the used shape analysis framework is the calculation of intrinsic means which are rate invariant. This helps to summarize the shape of trajectories belonging to the same class and accelerates the classification.

- *abnormal gait detection:* The inter-joints distances are used as features to detect the abnormal gait. These features have the advantage to be pose, position invariant and discriminative to model the human articulations movement in a given frame. The abnormal gait detection is performed on DAI gait dataset following the state-of-the-art protocol and show that the proposed approach success to classify abnormal and normal human actions. The result of experiments reports show that some distances related to the knees, ankles and the feet are more relevant than other distances.

- *multi-view 3D human object interaction dataset:* We collect a new multi-view 3D dataset for the purpose of providing an evaluative framework that supports analyzing abnormal and normal human activities with human object interactions. We evaluate the performance of our new multi-view dataset using the proposed feature by different scenarios: recognization from different views and synchronization of different views. The experiments on our multi-view 3D human object interaction dataset prove the effectiveness of our method.

1.3 Thesis organization

The rest of the paper is articulated as follow. In Chapter 2, we lay out related works in the area of human object interaction recognition, abnormal activity recognition as well as the issue of human behavior understanding for RGB-D data. In Chapter 3, we introduce the spatial modeling frame work that we employ to represent the human behavior and human object interaction. Chapter 4 presents the temporal modeling frame work to investigate rate invariant human object interaction. Chapter 5 presents the method that we employ for the task of human object interaction recognition and its evaluation in comparison with state-of-the-art on several benchmark action datasets. In Chapter 6, we adopt proposed method on another application: abnormal activity recognition and test our model on new collected multi-view 3D human object interaction dataset. Finally, we conclude this manuscript by summarizing the contributions of this thesis, enumerating remaining open problems and proposing directions for future research.

State-of-the-art 2

Sommaire

2.1 Introduction

Understanding human object interactions is critical for extracting meaning from everyday visual scenes and requires integrating complex relationships between human pose and object identity into a new perception. Analysis of human activities and behaviors through visual data has attracted a tremendous interest in the computer vision community. Indeed, this represents a task of interest for a wide spectrum of areas due to its huge potential, like human-machine interaction, physical rehabilitation, surveillance security, health care and social assistance, video games, etc [58].

In this chapter, we present the issue of object with daily human activities and relationship between human action and human object interaction. Then, we discuss the main applications where human object interaction and action recognition can be involved. Besides depth sensor technologies which able acquirement of depth images are introduced. Benchmark datasets of different kinds of data collected for the task of human object interaction analysis systems are presented. Finally, a review of the state-of-the-art approaches is presented and discussed.

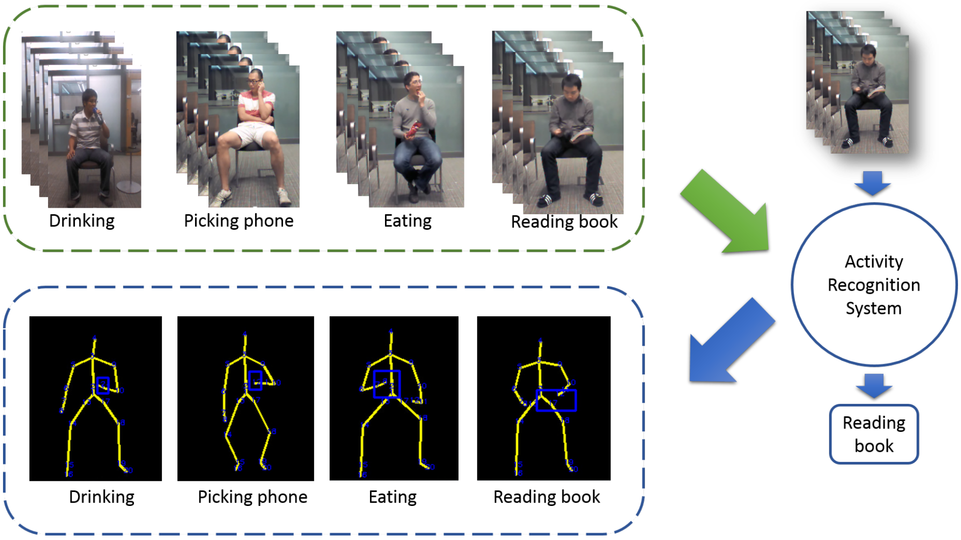


Figure 2.1 – *Illustration of human activity recognition system process. The input is an unknown activity sequence, the output is an labeled activity of this sequence.*

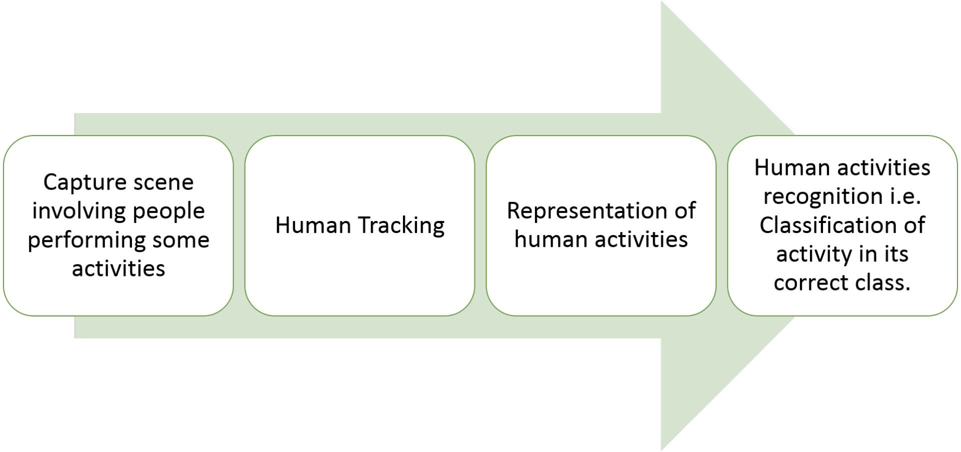

Capture scene involving people performing some activities
Human Tracking
Representation of human activities
Human activities recognition i.e. Classification of activity in its correct class.

Figure 2.2 – *General framework for human activity recognition system*

2.1.1 Human activities recognition system

A generic action or activity recognition system can be viewed as proceeding from a sequence of images to a higher-level interpretation as illustrated in Figure. 2.1. There are several annotated classes of activity sequences, and an unknown activity as observed test sequence which should be recognized by activity understanding system. The task of this system is identifying and understanding the activities occurring in the videos which consist of massive amounts of raw information in the form of spatio-temporal pixel intensity variations.

As shown in Figure. 2.2, a general overview of the components of a human activity recognition system. They include a scene capturing system, human tracking, activity representation methods, activity recognition methods, applications, and datasets. The detailed discussion is given in subsequent section.

2.1.2 Human activities terminology

A realistic depiction of human activities often involves complex interaction with the environment. So an initial definition of human behavior terminology is essential before discuss human object interaction in detail. There are various types of human activities [3], which are categorized into four different levels depending on their complexity as illustrated in Figure. 2.3 and defined as follows:

- Gestures: Gestures are defined as the elementary movements of a person's body part, and are the basic components describing the meaningful motion of a person. For instance, 'Stretching an arm' and 'raising a leg' are good examples of gestures.
- Actions: Actions are characterized by single person activities that may be composed of multiple gestures organized temporally without holding any objects, such as 'walking', 'waving', and 'punching'. In addition, like the example 'walking', can be performed in different cases: abnormal gait and normal gait. This exceptional case is easily found in daily life and also a significant issue for human activities analysis.
- Interactions: Interactions are highly semantic comprehension of human activities that involve two or more persons and/or objects. For example, 'picking phone' and 'using laptop' are interactions between one human and one object. Nevertheless, 'two persons fighting' is an interaction between two humans and 'a person stealing a suitcase from another' is a human-object interaction involving two humans and one object.
- Group Activities: Group activities are the activities performed by conceptual groups composed of multiple persons and/or objects. 'A group of persons marching', 'a group having a meeting', and 'two groups fighting' are typical examples of them.

Except the complexity of these different kinds of human activities, the execution time of them are also proportional to the complexity degree. During this thesis, we are more concern with human object interactions and distinguish between abnormal gait and normal gait, which is more close to real-world conditions and daily scenes. Hence, more research needs to be done to address these practical issues.

2.2 Applications

Vision based human motion recognition has fascinated many researchers due to its critical challenges and a variety of applications. The applications

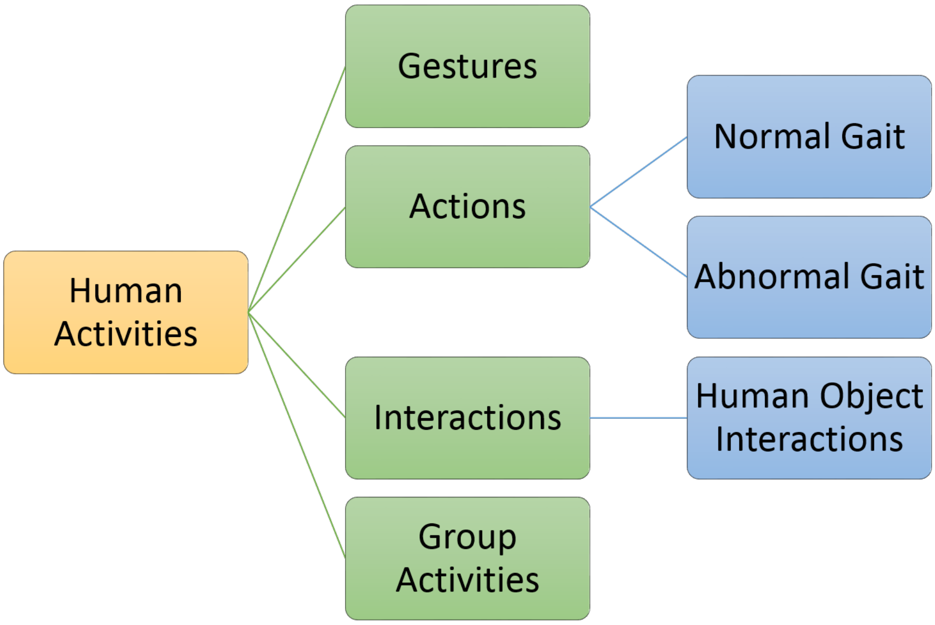


Figure 2.3 – *Taxonomy of human activities according to the complexity*

range from simple gesture recognition to complicated behavior in surveillance system. This leads to major development in the techniques related to human motion representation and recognition [53] [109].

The objective in an activity-driven application is to analyze and identify activities so that their semantic meaning can be understood in each specific domain, and approaches to construct human representations have been widely used in a variety of real-world applications, including video analysis [39], surveillance [51], robotics [29], human-machine interaction [101] augmented and virtual reality [26], assistive living [81], smart homes [15], education [72], and many others [17], [54], [36].

Here, we present some application areas that will highlight the potential impact of vision-based activity recognition system.

1. *Smart Surveillance:*

   Nowadays the surveillance systems are automatically tracking individuals so that human operators can view the video contents continuously. However, it is a tough task for human to monitor all the videos all the time, especially during period of the rapid development of RGB-D sensors. In reality, we need an intense requirement of smart surveillance systems to analysis the contents of video after a mishap. Not only this kind of systems can be online but also identify human behavior and motion accurately.

During more special situations and places like military territory, hospitals, traffic congestion analysis, distant human identification, abnormal behavior detection, schools, government buildings, commercial premises, and railway stations [66] [75], smart surveillance is required. Recently, the concept of smart home attracts a lot of attention from researchers in computer vision domain and will bring a big change in living condition and habitancy way [41].

2. *Behavioral Biometrics:*

   The study of biometrics is utilizing human physical or behavioral cues to recognize them. In past decades, finger print, face or iris can be classified as biometrics on account of relying on physical attributes to perform recognition task. Nowadays, the gait pattern as a biometric has gaining popularity. The main advantage of the recognition of the gait pattern is that subject cooperation is not necessary as compared to the other biometrics [95].

3. *Entertainment and Art:*

   Human motion synthesis finds wide use in the gaming and animation industry like Microsoft Xbox and the work of [52] that recognized sequences of dance movements from depth data. It facilitates to improve the quality of alteration of the movements and increase the effectiveness of a scene.

4. *Medical:*

   Human motion recognition is utilized in medical field for both of doctors and patients in many areas such as neurology, body posture, orthopaedics and fitness. For example, detection of abnormal behaviors like observing the performance disabled patients in their room is helpful for doctors to analysis possible deficiency in their motions and offer efficient treatment. Except that, elderly people may have physical or mental diseases can be well taken care of by intelligent systems designed based on human behavior analysis [76] [64].

5. *Gesture and Posture Recognition and Analysis:*

   One of the enduring challenges in human-computer interfaces is understanding the interaction between a human and a computer. Human gesture and posture recognition play an important role in advanced natural interface with computers and computerized systems. The promising applications such as smart room, sign language recognition, controlling devices, and others [93] [96] to bring about improvements in computers that can better interact with humans.

6. *Robotics:*

   In most recent technological developments, robots still have generally supported limited, scripted interactions, often relying on a human operator to help with input processing and appropriate behavior selection. So the behavior understanding is significant for robotic applications. For instance, the robotic designed activities can produce use scenarios to guide the development, increase user acceptance toward robotic applications in consumer markets and develop success metrics for human-interactive mobile robots (for entertainment or companion purposes), in particular those which targeted to untrained users for multiple assistive purposes.

7. *Sports and Exercise:*

   There is another study and practice of human activity recognition is sport which designed for analyzing athletic movements and efficient frameworks for training [11]. Nevertheless, monitoring system for the elderly people exercises [27] and feedback system for rehabilitation exercise proposed in [118].

2.3 Scene capturing and human tracking

The applications are key to the selection of a capturing system. Many applications are adopted multiple cameras or single camera. It is obvious that processing the contents captured by single camera system is easier than multiple camera system, but they may miss the detailed human features. Even though the multiple camera system handle this problem well, the procedure is complicated.

The initial elements of human activities analysis from video content is detecting subjects as well as corresponding objects. There are many constraints like clothing, illumination, background, resolution and frequency of frames need to be considered. As technology continue to evolve, advanced capture systems like *Mircosoft Kinect* [69] and *ASUS Xtion Pro Live* [8] emerged. These kind of systems predigest capture scenes procedures and human tracking. Especially, these systems provide us human skeleton joint data, depth data of objects and so on. The obtained data can be applied on representation methodologies and recognition algorithms for human activities recognition.

Human tracking is the process of identifying human body pose from videos. There are many challenges like partial occlusion and variations in lighting conditions for estimation of human body pose tracking. Due to the several advantages of these data: easy to remove background; isolating and tracking human body; allowing capturing the human motion at each frame. Kinect has changed the picture by providing its depth data and skeleton stream as an output for further processing on video-based human activities recognition.

2.3.1 A brief review of Kinect

Microsoft released the software package as well as the hardware of Kinect for Windows V1 [69] in November 2010, Kinect for Windows V2 [70] in July 2014 which price is at the consumer level for domestic use. There are two versions of Kinect shown in Fig. 2.4. The hardware configuration of Kinect generally consists of an RGB camera and an depth sensor which provide color images and depth images respectively. The software tools of Kinect include official SDK which is Kinect Windows SDK [71] and unofficial SDK, e.g., OpenNI [82], which provide a straightforward access to RGB and depth data. The Kinect Windows SDK supports multiple

Kinect sensors and provides a whole body tracker. But we note that the majority of RGB-D datasets of human activities recognition being use are generated with Kinect V1, only a few datasets created by Kinect V2.

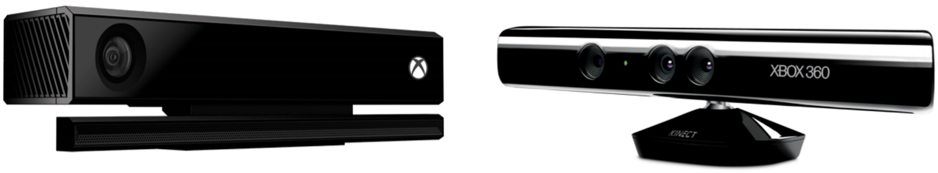
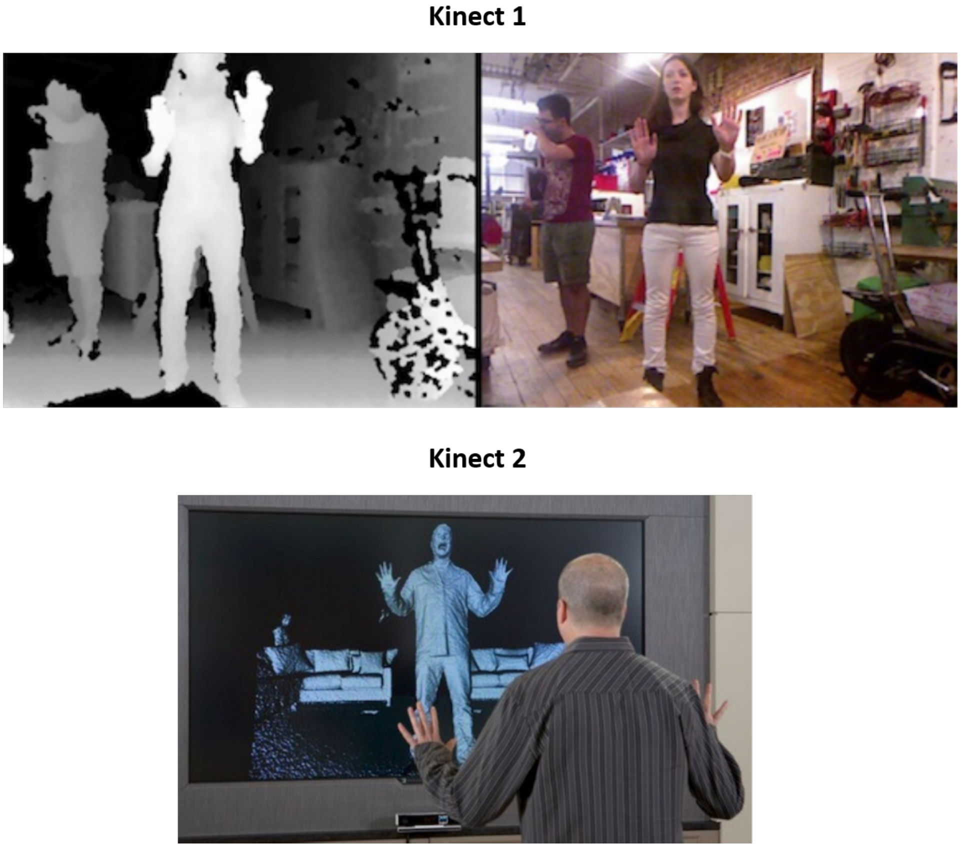


Figure 2.4 – *Illustrations of Kinect sensor. Left one is Kinect2; Right one is Kinect1.*



Kinect 1

Kinect 2

Figure 2.5 – *Illustration of resolution capability of Kinect1 and Kinect2.*

Kinect for Windows V1 is structured light sensor which is different from other RGB cameras like stereo cameras [107] which are sensitive to light changes, a time-of-light (ToF) cameras [129] which have high speed, depth image covering every pixel and high price. But Kinect for Windows V2 adopts time of light sensor providing higher resolution capability. The comparison 2.5 represents the visual input difference between the Kinect V1 and Kinect V2.

These advantages of Kinect for Windows sensor and SDK make them extend to computer vision field and suitable for many applications, such as 3D-simultaneous localization and mapping (SLAM) [47], people tracking [80], object recognition [13], and human activity analysis [25] [65] and so on.

2.3.2 RGB-D Data

As sensors continue to evolve, the new descriptor is about 3D positions of human body joints appears. The following graphics show the skeletal joints the Kinect V1 and Kinect V2 return.

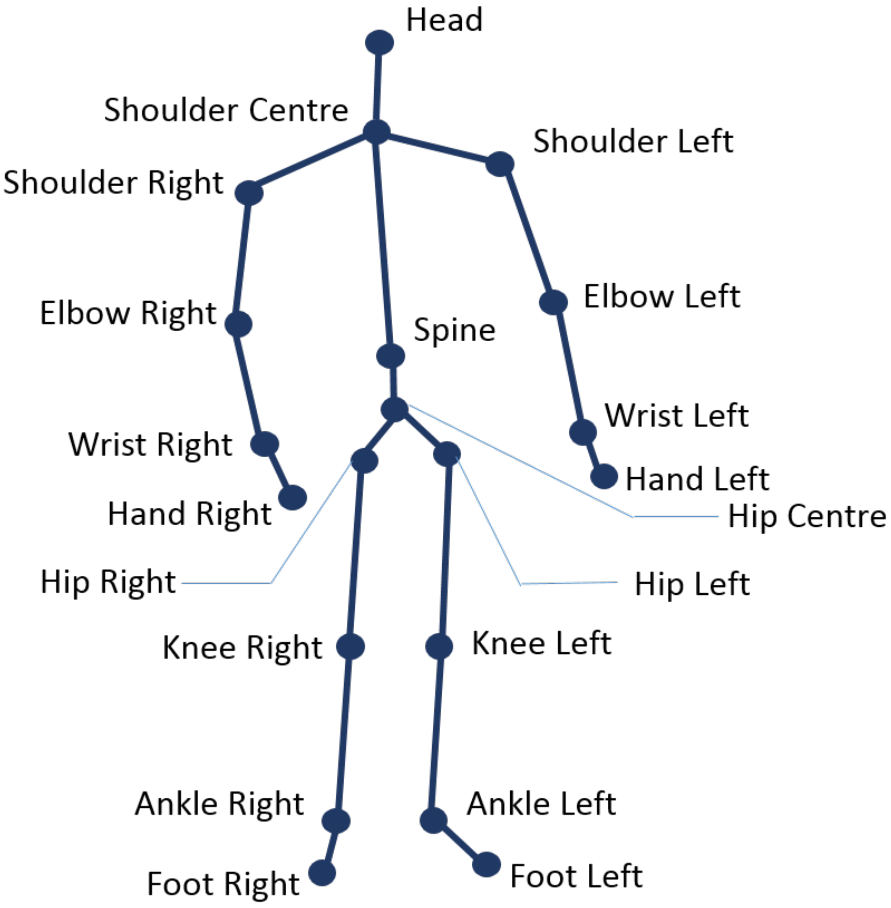


Figure 2.6 – *Kinect v1 ID joint map*

Thanks to the seminal work of [98], it facilitates the research in human activities recognition by estimating the joint locations of a human body

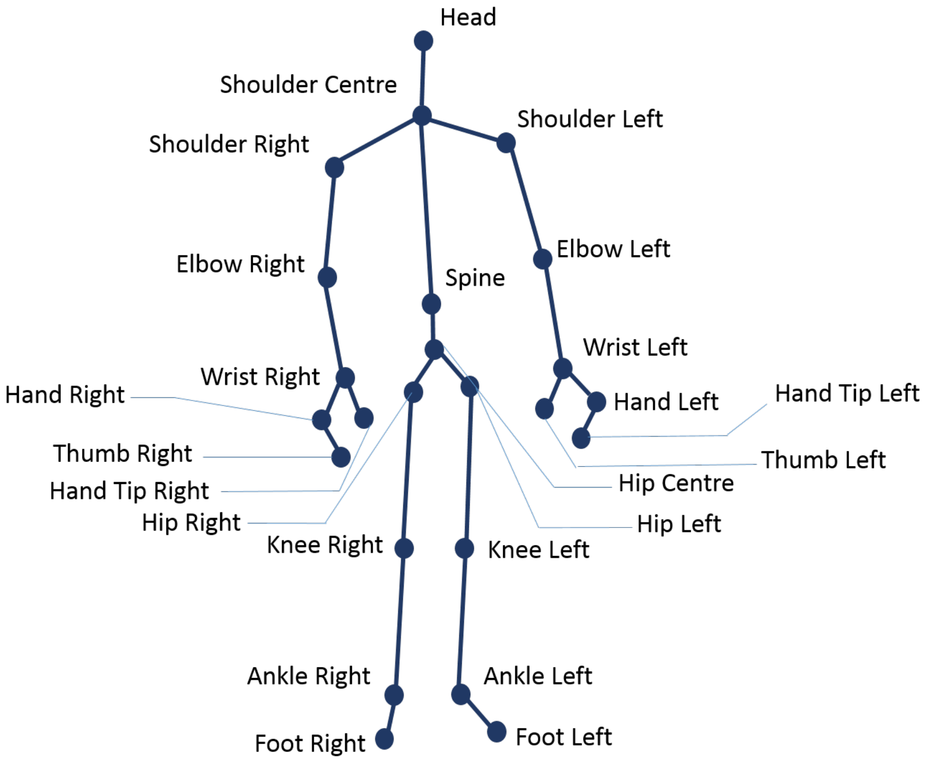


Figure 2.7 – *Kinect v2 ID joint map*

from depth map accurately. Due to depth map being proved to provide the data for an efficient human body estimation, human pose can be represented as a set of 3D humanoid skeleton joints. Kinect V1 was able to estimate 20 skeleton joints of the whole body and Kinect V2 increased to 25 joints with more detailed information around two hands.

2.4 Datasets

Analysis of human activities and behavior through visual data has attracted a tremendous interest in the computer vision community [58]. As the emergence of 3D data reduces the challenges to human behavior analysis, several datasets have been collected to serve as benchmark for researchers algorithms. The datasets are categorized into three classes: single-view activity, multi-view activity and multi-person activity. If only from literally, the action or activity is of single-view activity datasets are captured by only one specific view point. For multi-view activity datasets, there are two or more view points of each action or activity being captured.

Behind the literal meaning, we should note that each action or activity is performed by one actor at a time. In addition, the multi-person activity datasets are composed of not only interactions between two people but also activities performed by multiple persons.

In thesis, we focus our study on human object interaction recognition so adopt MSRDaily Activity 3D dataset [116] and 3D Online Action dataset [139]. We also evaluate our work for abnormal gait detection on DAI gait dataset [21]. Due to few numbers of abnormal and normal gait datasets, there is no survey for abnormal and normal gait datasets. We will introduce DAI gait dataset in details in the later chapter.

Here, we will review the representative RGB-D datasets based on the three categories and created by Kinect sensor about human activity analysis [142] [19]. And also the datasets contain human object interaction are shown in the table. In Table 2.1, we list the characteristics of the selected RGB-D datasets.

MSR Action 3D dataset

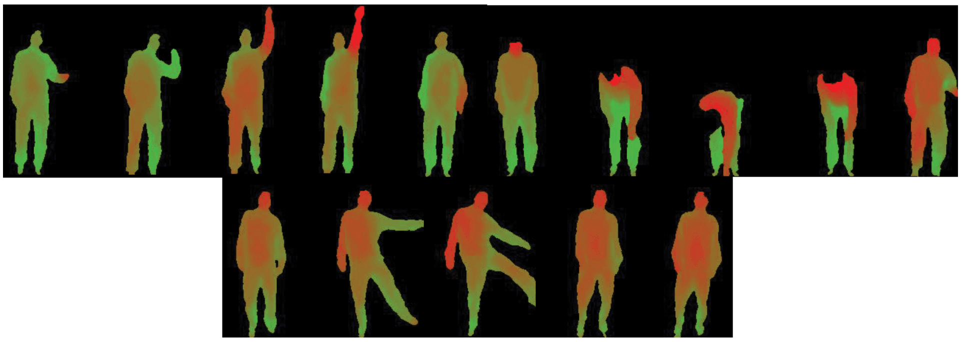


Figure 2.8 – *Samples from MSR Action3D dataset [116]*

MSR Action3D dataset [116] is the first public benchmark RGB-D action dataset collected by Microsoft Research Redmond and University of Wollongong in 2010. The dataset contains 20 actions: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis serve, golf swing, pickup and throw*. Each action was performed by ten subjects for three times. In this dataset, all of actions performed by each subject facing the camera at a fixed point. Note that

Table 2.1 – *The characteristics of the selected RGB-D datasets*

No.	Dataset	Year	Data	Size	Context	Composable interactions
1	MSR Action3D [116]	2012	Color, depth, skeleton	10 subjects, 20 classes	single-view	No
2	MSRDailyActivity3D [116]	2012	Color, depth, skeleton	10 subjects, 16 classes	single-view	Yes
3	Cornell Activity CAD-60 [105]	2011	Color, depth, skeleton	4 subjects, 12 classes	single-view	Yes
4	Cornell Activity CAD-120 [60]	2013	Color, depth, skeleton	4 subjects, 10 classes	single view	Yes
5	Daily activities with occlusions [1]	2015	Color, depth, skeleton	1 subjects, 12 classes	single-view	No
6	DGait [14]	2012	Color, depth	55 subjects, 11 classes	single-view	No
7	3D Online action [139]	2012	Color, depth, skeleton	36 subjects, 7 classes	single-view	Yes
8	RGBD-HuDaAct [78]	2013	Color, depth	30 subjects, 12 classes	single-view	Yes
9	TST fall detection [38]	2014	Color, depth, skeleton	11 subjects	single-view	Yes
10	UR fall detection [61]	2014	Color, depth	30 falls, 40 activities	multi-view	Yes
11	UTD-MHAD [23]	2015	Color, depth, skeleton	8 subjects, 27 classes	single-view	No
12	Workout SU-10 exercise [77]	2013	Color, depth, skeleton	12 subjects, 10 classes	single-view	No
13	Composable activities [63]	2014	Color, depth, skeleton	14 subjects, 16 classes	single-view	Yes
14	MSRActionPairs [83]	2013	Color, depth, skeleton	10 subjects, 12 classes	single-view	Yes
15	MAD [49]	2014	Color, depth, skeleton	20 subjects, 35 classes	single-view	No
16	Cocurrent Action [121]	2013	Skeleton	12 classes	single-view	Yes
17	UTKinect [127]	2012	Color, depth, skeleton	10 subjects, 10 classes	single-view	No
18	UFCKinect [35]	2013	Skeleton	16 subjects, 16 classes	single-view	No
19	UPCV [106]	2014	Skeleton	20 subjects, 10 classes	single-view	No
20	SYSU [48]	2015	Color, depth, skeleton	40 subjects, 12 classes	single-view	Yes
21	RGB-D activity [123]	2015	Color, depth	7 subjects, 21 classes	single-view	Yes
22	G3DI [12]	2014	Color, depth, skeleton	12 subjects, 6 classes	multi-person	Yes
23	Multiview 3D event [120]	2013	Color, depth, skeleton	8 subjects, 8 classes	multi-view	Yes
24	UWA3D Multiview [92]	2014	Depth	10 subjects, 30 classes	multi-view	No
25	SBU [140]	2012	Color, depth, skeleton	7 subjects, 8 classes	multi-person	Yes
26	LIRIS [122]	2014	Color, depth	21 subjects, 10 classes	multi-person	Yes

the subject was required to perform an action by right arm or leg when this action only needed one are or leg.

Cornell Activity CAD-120 dataset

Cornell Activity CAD-120 dataset [60] contains activities and object interactions and is collected by the Cornell University. It includes 10 types activities performed by 4 subjects, and each activity was performed twice with different objects. These activities are *making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, picking objects, cleaning objects, taking food, arranging objects, having a meal*. Note that there are some activities performed by same subject with different objects. The data of this dataset consists of color image, depth image and skeleton data. Here, its skeleton only has 15 joints. So CAD-120 dataset is applied on human activity analysis and also object detection.

Figure 2.9 – *Samples from Cornell Activity CAD-120 [60]*

RGB-D activity dataset

RGB-D activity dataset [123] was recorded by the Kinect V2 camera and collected by Cornell University and Stanford University in 2015. It contains interactions with different objects in each video. As Kinect V2 has higher resolution of RGB-D data, so the body tracking was improved that skeleton data consists of 25 body joints. Here, 7 subjects performed 21 type activities which were in different environment: 10 in the office, 11 in the kitchen. These 21 activities interacted with 23 types of objects: *turn-on-monitor, turn-off-monitor, walking, play-computer, reading, fetch-book, put-back-*

book, take-item, put-down-item, leave-office, fetch-from-fridge, put-back-to-fridge, prepare-food, microwaving, fetch-from-oven, pouring, drinking, leave-kitchen, move-kettle, fill-kettle, and plug-in-kettle. Except complex background, the activities were performed relative to different views.

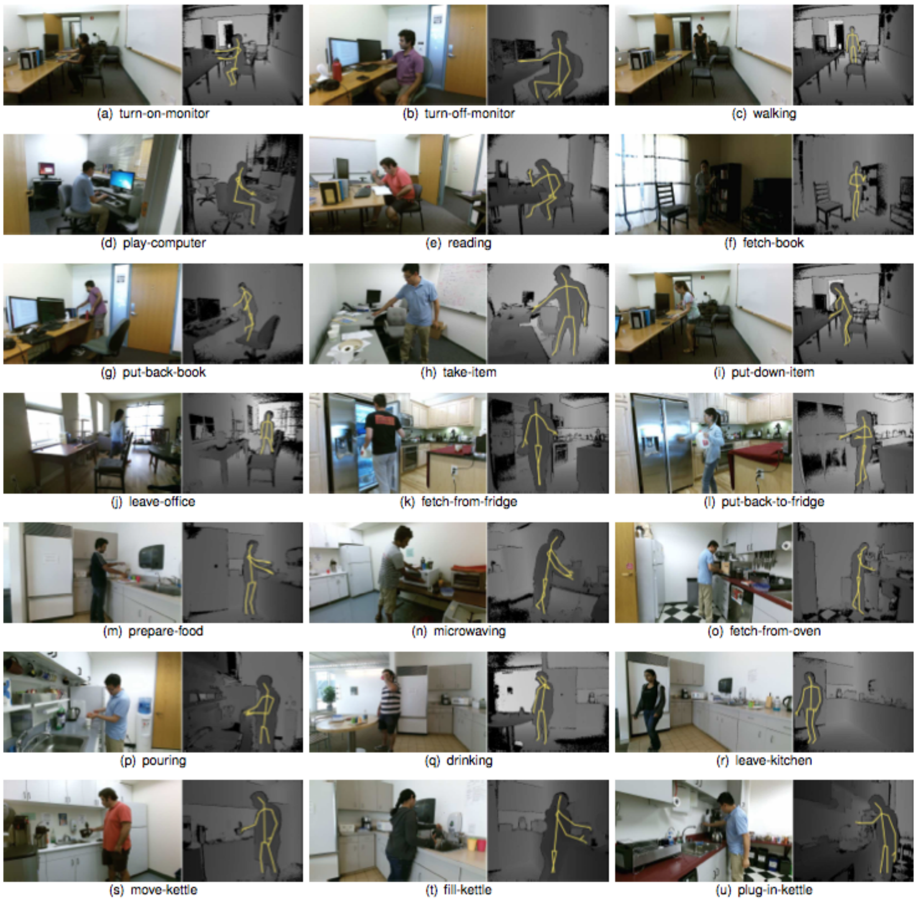


Figure 2.10 – *Example color, depth and skeleton frames from RGB-D activity dataset [123]*

RGBD-HuDaAct dataset

RGBD-HuDaAct dataset [78] as collected by Advanced Digital Sciences Center Singapore in 2011. It contains 12 categories of human daily activities performed by 30 subjects and motivated by the definitions provided by health-care professionals, including: *make a phone call, mop the floor, enter the room, exit the room, go to bed, get up, eat meal, drink water, sit down, stand up, take off the jacket and put on the jacket*. The subjects performed 2-4 repetitions of each action. In this dataset, there are human object interactions and no restriction on which leg or hand was used.

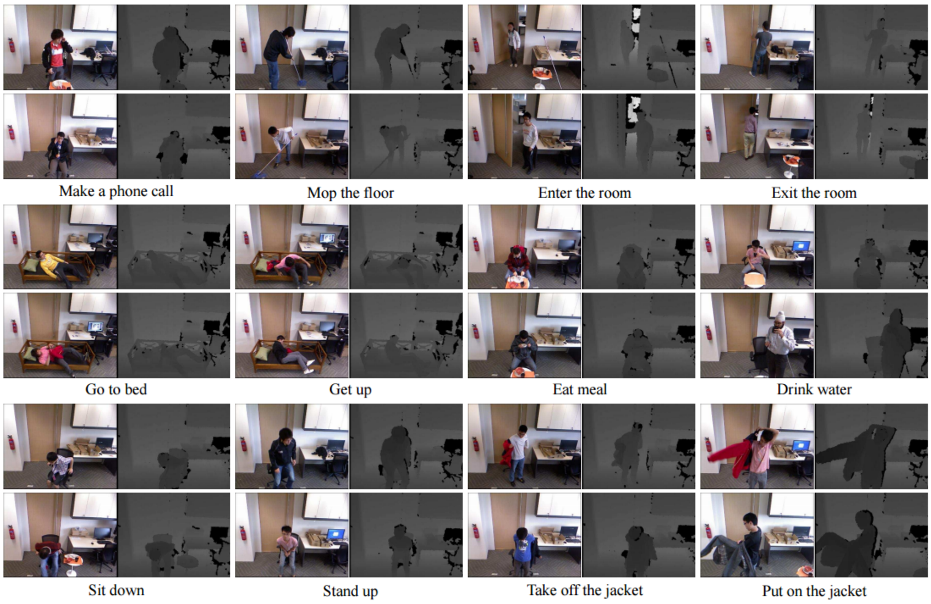


Figure 2.11 – *Example color and depth frames from RGBD-HuDaAct dataset [78]*

G3Di

G3Di [12] is a human interaction dataset for scenarios about multi-player gaming and was collected by Kingston University in 2014. Its data include color image, depth map and skeleton data. Specially, the dataset adopted a game sourcing approach where the users were recorded whilst playing computer games. This dataset contains 12 subjects split into 6 pairs. Each pair interacted through a gaming interface showcasing six sports involving several actions: *boxing (right punch, left punch, defend), volleyball (serve, overhand hit, underhand hit, and jump hit), foot- ball (kick, block and save), table tennis (serve, forehand hit and backhand hit), sprint (run) and hurdles (run and jump)*. In addition, most sequences were captured by a fixed camera containing multiple action classes in a controlled indoor environment.

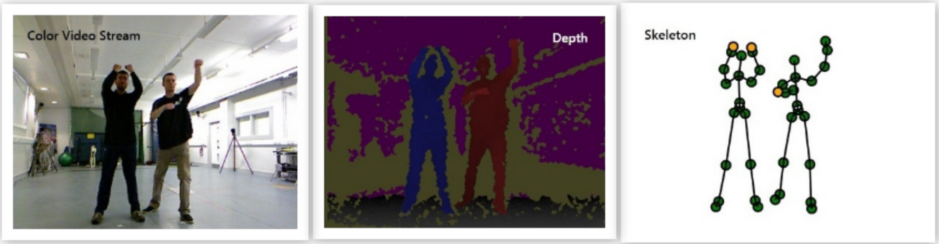


Figure 2.12 – *Example color, deptnbh and skeleton frame from G3Di dataset [12]*

Multiview 3D event dataset

Multiview 3D event dataset [120] is a large-scale multi-view 3D dataset created by University of California at Los Angles in 2013. The dataset obtained by utilizing three stationary Kinect cameras simultaneously at different viewpoints around the subjects. It contains three kinds of data: RGB, depth and skeleton. This dataset contains 8 classes of events performed by 8 subjects 20 times independently with different object instances and in various styles. The 8 event classes are: *drink with mug, call with cellphone, read book, use mouse, type on keyboard, fetch water from dispenser, pour water from kettle, and press button*. These events involve 11 object classes: *mug, cellphone, book, mouse, keyboard, dispenser, kettle, button, monitor, chair, and desk*. There are human object interactions in this dataset.

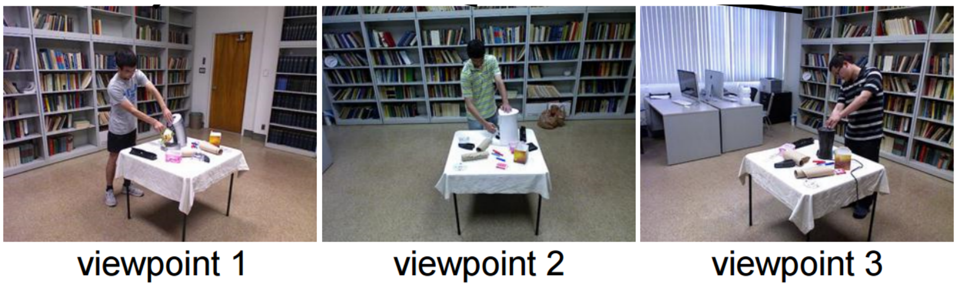


Figure 2.13 – *Samples from Multiview 3D event dataset [120]*

2.5 Related Work Concerning Human object interaction

In the literature of activity recognition, many previous work in behavior analysis used videos produced by conventional cameras [111], [102], [2], [68].

Recently, with the development of the commodity depth sensors like Kinect, there has been a lot of interests in human action recognition from depth data such as [130], [108], [99], [7], [128], [24], [50], [87]. Instead of covering all ideas exhaustively, we direct interested readers to some recent surveys [3], [22], [56], [145] that together overview this domain.

In this section, we present and summarize previous papers on the recognition of human object interaction and the work of action recognition using human body descriptors which most closely related to our approach. Additionally, we also apply our proposed algorithm on abnormal gait detection. So there is a brief survey about abnormal gait detection. In this section, we briefly review related work about human object interaction and action recognition from four streams of research: 2D information, 3D skeleton information, depth information and hybrid information. Next, we discuss the work based on these four streams respectively.

2.5.1 Methods based on 2D information

There are a large amount of existing methods for human-object interaction recognition based on static and 2D videos such as [88] [68] [28] [124] [57] [4] [73] [144]. Besides videos, some works were based on 2D images. For example, [42] combined spatial and functional constraints between human and objects to recognize actions and objects on static images. [31] [30] learns a discriminative deformable part model (DPM) that estimates both human poses and object location.

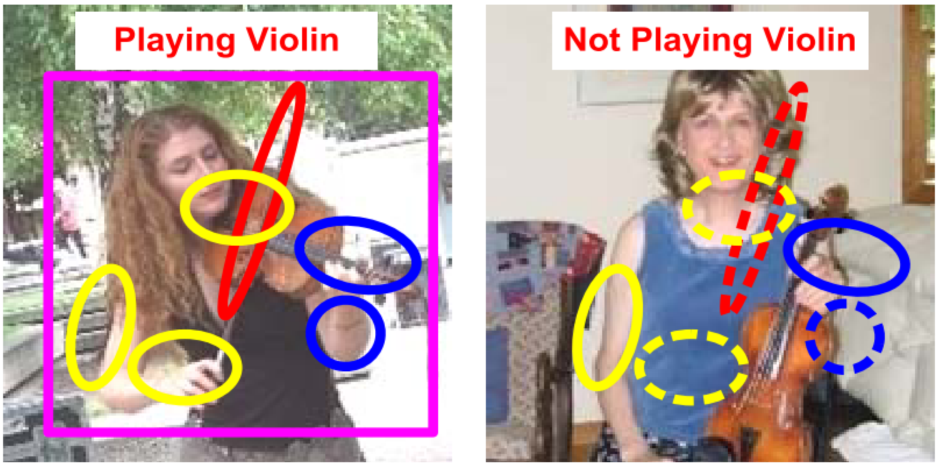
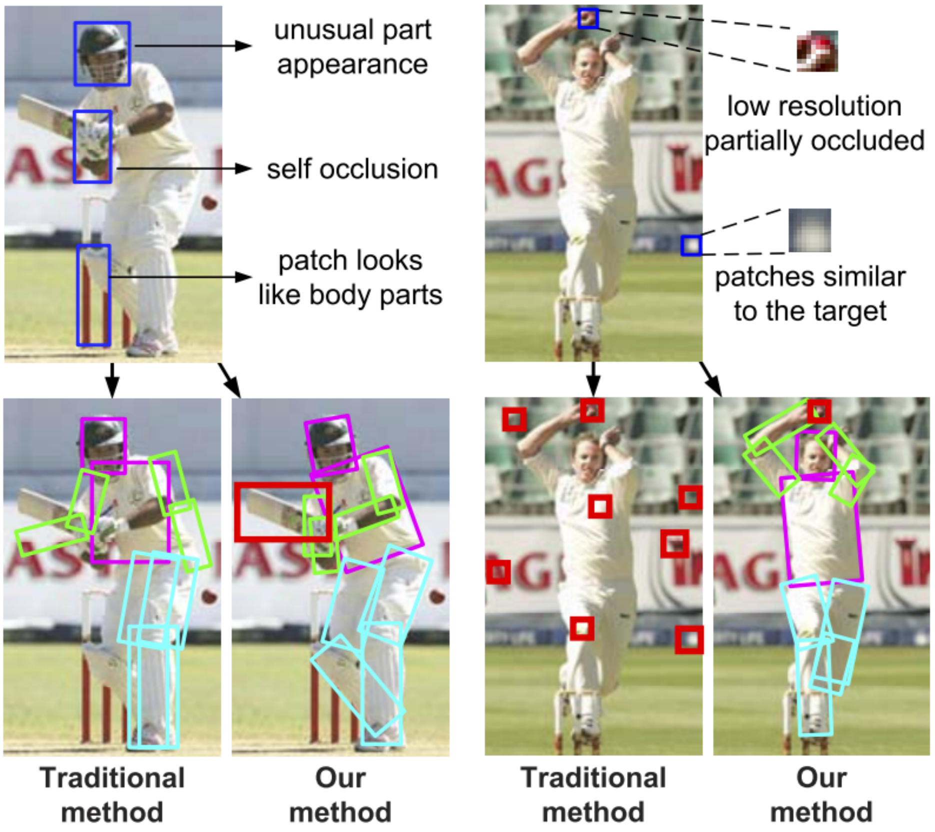


Figure 2.14 – *Example of grouplets representation [136]*

As introduced above, they have made the object recognition and motion estimation independent. On the other hand, some researchers have studied to design a model representing mutual information between objects and human actions. Here, we present the representative works for such task. [136] adopted grouplet encode detailed and structured information from the images to estimate the 2D poses. Their algorithm discovered the features called grouplets to describe the mutual information between humans and objects by encoding the position, appearance and shape of images patches. As shown in the left Figure. 2.14, three sample grouplets are shown in three different colors.

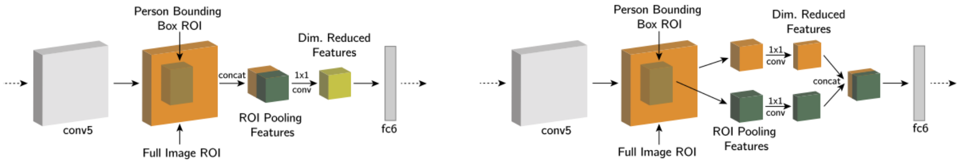
Similar to [136], [137] proposed a model to exploit the mutual context of human poses and objects in one coherent framework. They treated object and human pose as the context of each other in human-object interaction activities. We can see from Figure. 2.15, the object was detected for better understanding the human object interaction.



(a) Human pose estimation (b) Object (ball) detection

Figure 2.15 – *Objects and human poses can serve as mutual context to facilitate the recognition of each other. [137]*

[89] also inferred the spatial information of objects by modeling the 2D geometric relations between human body and objects. Afterwards, the spatio-temporal feature was adopted such as [135] [134]. They developed spatio-temporal AND-OR graph to model the spatio-temporal structure of the pose in an action.



(a) Fusion-1 (b) Fusion-2

Figure 2.16 – *The illustration of network architecture from [67]*

For fine-grained human object interaction recognition, [143] used the MSRDaily Activity 3D dataset obtained by Kinect and linked object proposals followed by feature pooling in selected regions. In their work, the proposed method only analyzed 2D video content without depth map. But they added skeleton information to localize useful interaction parts and remove background noise.

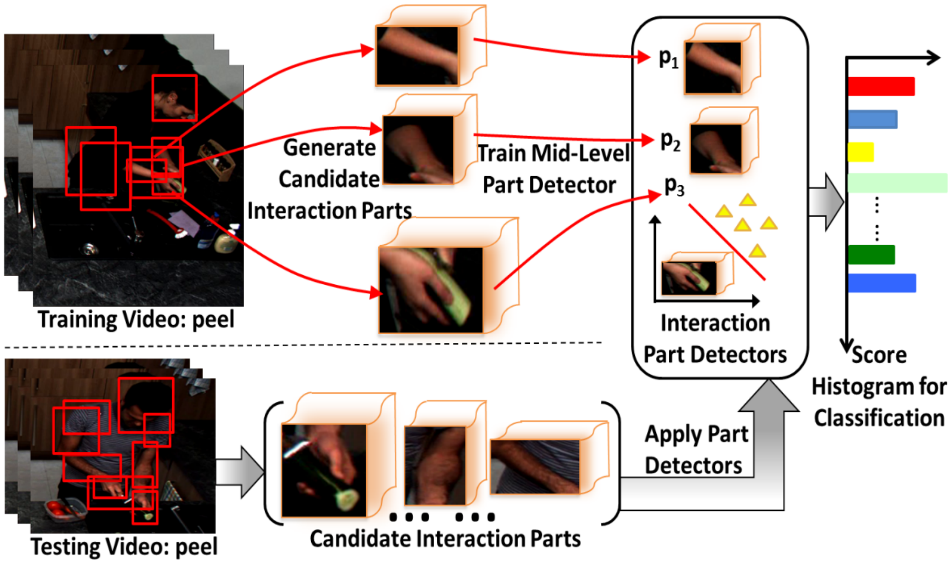

Training Video: peel
Generate Candidate Interaction Parts
Train Mid-Level Part Detector
p_1
p_2
p_3
Interaction Part Detectors
Score Histogram for Classification
Testing Video: peel
Candidate Interaction Parts
Apply Part Detectors

Figure 2.17 – *The illustration of the approach for fine-grained human object interaction recognition [143]*

Recently, [67] proposed a simple deep convolutional neural networks (CNNs) models. Then, they fused the features from person-bounding boxes, global image context and person appearance to detect human activity labels. The multiple instance learning framework was used to predict human object interactions and the network architecture about this work is shown in Figure. 2.16.

[9] proposed a framework for achieving the centerpiece interaction recognition by capturing the interesting objects of an image. They utilized

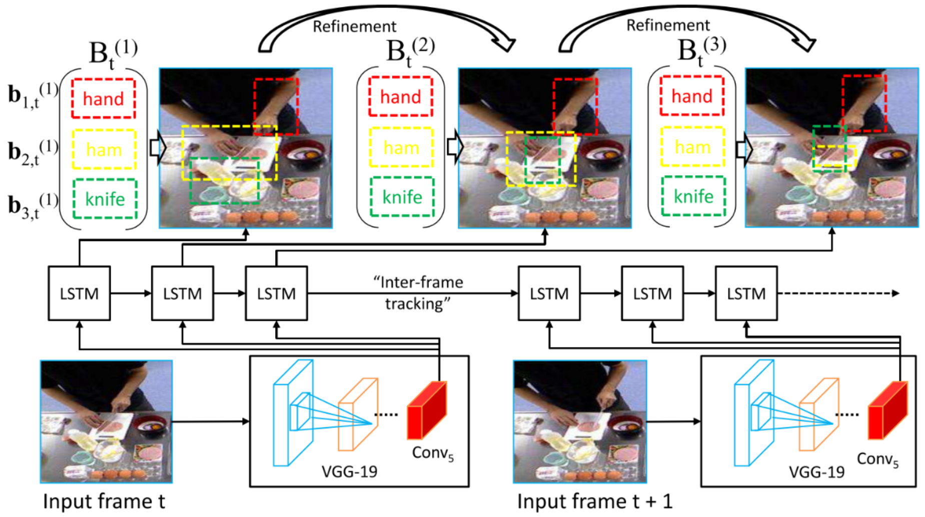


Figure 2.18 – *Overview of the progressive interactional object parsing method based on LSTM network. [79]*

2.5D spatial co-occurrence context among objects and designed a hierarchical model to learn the features of objects. In [79], a progressive interactional object parsing method was proposed. This method adopted recurrent neural network to implement interactional object parsing for each frame. Note that they used a set of long-short term memory (LSTM) nodes instead of all object detection frame by frame, which contained more information to local objects. Then the results of object parsing were used for representing human object interaction.

These methods define the human-object interactions on 2D image. Such contextual cues are often compromised by the viewpoint changes and occlusions.

2.5.2 Methods based on depth information

Thanks to the work of [98] by using the depth cameras which offers a cost-effective method to track 3D human poses, many approaches in the literature adopted skeleton, RGB and depth these feature to model human-object interaction and human activities. In this part, we mainly discuss the works based on depth sequences.

Before the emergence of Kinect, there already were some research investigating on human object interaction and action recognition using depth maps obtained decent performance. [62] employed an action graph to model the dynamics of the actions and sample a bag of 3D points from the depth map to characterize a set of salient postures that correspond to the nodes in the action graph. But there are limitations of this work such as noise and occlusions in the depth maps and sampling scheme is view dependent. [37] extracted the objects from depth data captured from TOF camera to perform sub-activity (referred to as action) classification and functional categorization of objects. Their method first detected the sub-activity being performed using the estimated human pose from depth data, and then performed object localization and clustering of the objects into functional categories based on the detected sub-activity.

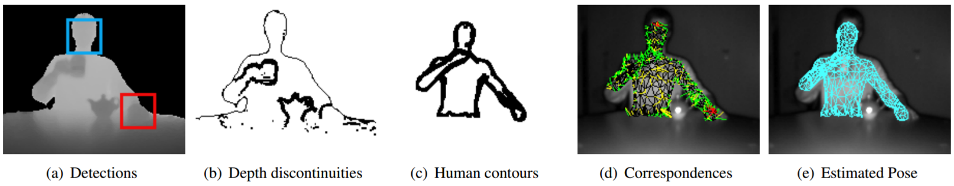

(a) Detections (b) Depth discontinuities (c) Human contours (d) Correspondences (e) Estimated Pose

Figure 2.19 – *Pose estimation from depth data. [37]*

[84] took advantage of pose tracks and depth readings and employed the latent structural SVM to train the model with part-based pose trajectories and object manipulations. In [83], they presented a descriptor histogram of oriented 4d surface normals (HON4D) capturing the distribution of the surface normal orientation in the 4d volume of time, depth and spatial coordinates from depth maps.

The idea of projecting each depth map onto these three orthogonal planes is also employed in [132]. Then, the motion energy is obtained by computing and thresholding the difference between two successive maps. Such motion energy is then stacked through all the video sequence resulting in a Depth Motion Map (DMM) for each view. Such DMM highlights areas where main motion takes place during the action. Histogram of Oriented Gradients (HOG) is then applied to DMM maps to extract features for each view. The three HOG features are concatenated to build a single feature for each sequence. A SVM classifier is trained on this feature to perform action recognition. Similarly to [62], this methods suffers from its view dependency.

Other works propose to extend the idea of spatio-temporal interest points (STIPs) to depth data. Indeed, its capability to handle clutter background and partial occlusions has been proven in RGB video. Hence, the work in [126] proposes to apply this idea to depth maps by extracting STIPs from depth video called DSTIPs. Then, the 3D cuboid around each DSTIP is considered to compute the depth cuboid similarity feature (DCSF) describing the local depth appearance within each cuboid. Finally the bag-of-words approach is employed to identify a cuboid codebook and represent the actions. In [133], a novel formulation of the cuboid descriptor is proposed based on its sparsification and its quantization. This feature called 3D sparse quantization (3DSQ) is then employed in a spatial temporal pyramid (STP) [94] for hierarchically describing the action. A similar idea of STIP is proposed by Rahmani et al. [46], where key-points are detected from the 3D point cloud. Each point is then described using the Histogram of Principal Components (HOPC). The main advantage of this method is its robustness to viewpoint, scale and speed variations.

Such robustness are important challenges investigated by many researchers. For instance, a binary depth feature called range-sample depth feature is proposed by Lu and Tang [20]. This feature describing both shape geometry and motion is robust to occlusion as well as possible changes in scale, viewpoint and background.

Instead of directly working on depth maps, other methods propose to consider a depth sequence as a 4D space (3D+t) divided into spatio-temporal boxes to extract features representing the depth appearance in each box. For instance,Vieira et al. [114] propose to divide the 4D space into a grid containing a certain number of 4D cells. Then the spatio-temporal occupancy pattern (STOP) is computed within each 4D cell. Such feature counts the number of point that fall into the spatio-temporal grid.

By applying a threshold on this feature, they are able to detect which 4D cells correspond to motionless (red in the figure) and which correspond to motion (green in the figure). The concatenation of such feature of each 4D cell is used to represent the depth sequences. A similar occupancy feature called random occupancy pattern (ROP) is also employed in

[116]. Differently, the 4D sub-volumes are extracted randomly at different locations and with different sizes.

The 4D space is also investigated in [83]. Then, the depth sequence is partitionned into spatio-temporal cells. Within each cell the orientation of 4D normals are quantified using 4D projectors to build a 4D histogram. This feature, called histogram of oriented 4D normals (HON4D), captures the distribution of the normal vectors for each cell. The idea of computing surface normals within spatio-temporal cells is also used by Yang and Tian [131] to describe both local motion and shape information characterizing human action. However, a limitation of such normal methods is that they assume correspondence between cells across the sequence. Hence, these methods may fail when the subject significantly changes his spatial position during the performance of the action.

In most cases, the works based on depth images adopt the whole depth maps which is difficult to achieve the real-time recognition. Skeleton descriptor can effectively solve this problem. So we take advantage of this property of skeleton information applied in our approaches.

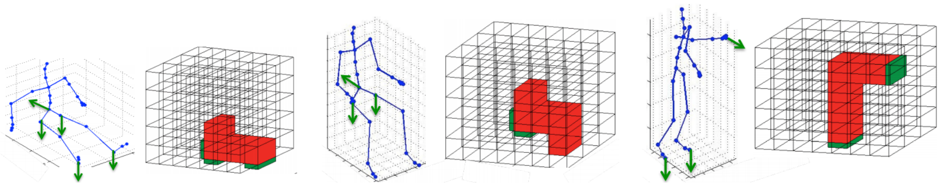
The depth data is not suitable for online action recognition from unsegmented streams.

2.5.3 Methods based on 3D skeleton information

In our work, skeleton information is the primary feature for modeling human object interactions. But as far as we know, most existing methods based on skeleton information are human action recognition and there are few works about human object interactions due to this complexity. Here, we will introduce the work about both of interactions and action recognition research based on skeleton data.

In early stage, [43] created a joint space can then be used to predict potential human poses and joint locations from a single image. This joint space modeled he physical interaction between human poses and scene geometry.

[40] performed the Dynamic Time Warping (DTW) on feature vectors defined by 3d joint trajectories. Here, their skeleton information was ob-



(a) Sitting Reclined (b) Sitting Upright (c) Reaching

Figure 2.20 – *Illustration of Qualitative Representation of Human Poses [43]*

tained from motion capture data. A transfer learning framework analyzed the joint-3D-trajectories on a spatio-temporal manifold. After that, Dynamic Manifold Warping (DMW) was introduced to align two human motion sequences not only 3D but also 2D trajectories and provide a motion similarity score.

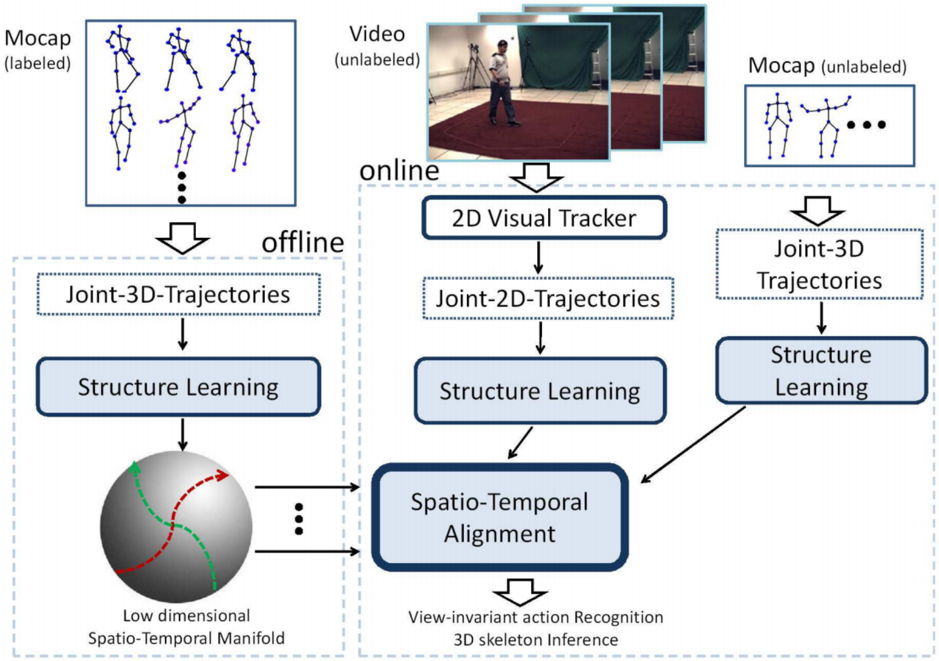


Figure 2.21 – *The flow chart of proposed method [40]*

As Kinect sensors make available a representative 3D humanoid skeleton frame by frame, [127] proposed an approach for human action recognition with histograms of 3D joint locations (HOJ3D) as a compact representation of postures is proposed. The HOJ3D computed from the action depth sequences are re-projected using LDA and then clustered into several posture visual words, which represent the prototypical poses of actions. The temporal evolutions of those visual words are modeled by discrete hidden Markov models (HMMs). [112] represented a human skeleton as a point in the Lie group which is curved manifold, by explicitly modeling the 3D geometric relationships between various body parts using rotations and translations. Using the proposed skeletal representation, it modeled human actions as curves in this Lie group then mapped all the curves to its Lie algebra, which is a vector space, and performed temporal modeling and classification in the Lie algebra.

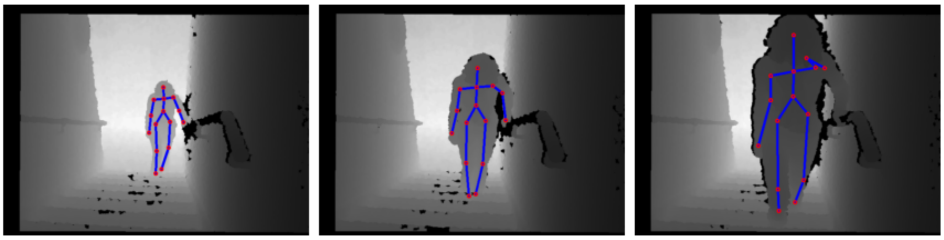


Figure 2.22 – *Illustration of skeletons at bottom, middle, and top of the stairs. Only the skeleton was used in this framework. [85]*

Later, the research put effort on online recognition system which skeleton information shows the big advantage of. The method [85] analyzed the quality of movements from skeleton representations of the human body. They used a non-linear manifold learning to reduce the dimensions of the noisy skeleton data. Then building a statistical model of normal movement from healthy subjects, and computing the level of matching of new observations with this model on a frame-by-frame basis following Markovian assumptions.

[139] proposed a middle level representation called orderlet by encoding the spatial configuration of a group of skeleton joints for real time recognizing human object interactions. This orderlet captured ordinal information that represented the distance variation between different pair of joints for modeling interactions. In their method, they also adopted object shape information extracted for depth images by LOP [117] to accomplish this task. Additionally, they collected a dataset named only including human object interactions for cross-environment and online action recognition. We also performed our approached by using depth and skeleton information on this dataset and details will be shown in the following sections.

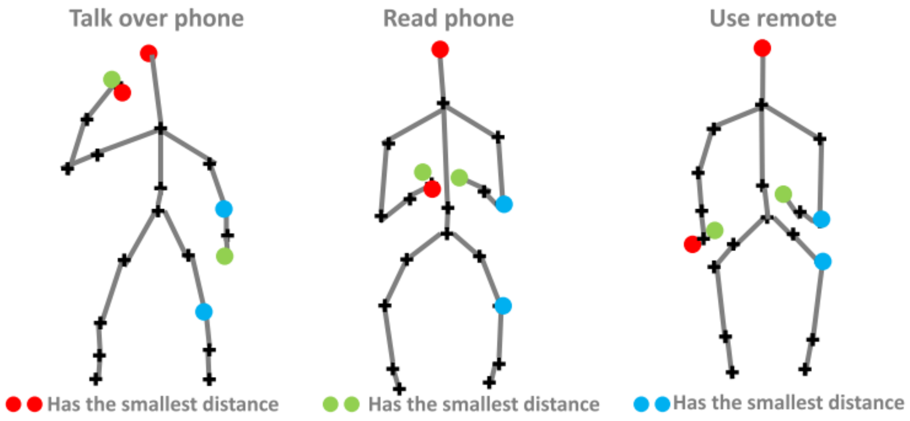

Talk over phone
Read phone
Use remote
●● Has the smallest distance
●● Has the smallest distance
●● Has the smallest distance

Figure 2.23 – *An illustration of our orderlet representation. [139]*

Recently, more works put effort on analyzing human object interactions based on skeleton information. [6] proposed to model the evolution of human skeleton shapes as trajectories on Kendall's shape manifolds, and used a parameterization-invariant metric [104] for aligning, comparing, and modeling skeleton joint trajectories, which can deal with noise caused by large variability of execution rates within and across humans. [90] learned dictionaries of sparse codes of sampled spatial-temporal 3D volumes from depth maps and achieved real-time human action recognition.

In the work of [115], their method only uses 3D skeletons and performed well on benchmark datasets like MSRDaily Activity dataset. They proposed an approach by mining a set of key-pose-motifs for each action class. These key-pose-motifs were figured by soft-quantization dictionary to improve accuracy. Last, the sequences were classified by matching them to the motifs of each class and selected by the maximum mathing score. [18] also tested their algorithm on MSRDaily Activity dataset and also only using skeleton representation.

[34] utilized the skeletal joints location as input for the task of segmentation, classification and prediction of ongoing human actions. They incorporated spatial and temporal characteristic of actions for such task. The also tested their framework on 3D online action dataset which we adopted in our work.

There are several works [33] [32] which introduced a human represen-

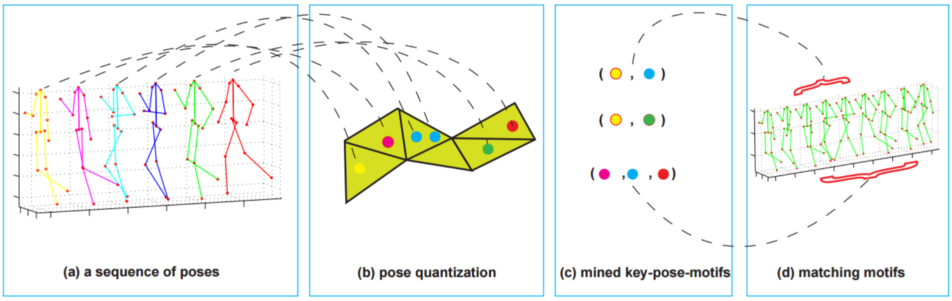


Figure 2.24 – *Overview of the method of [115]*

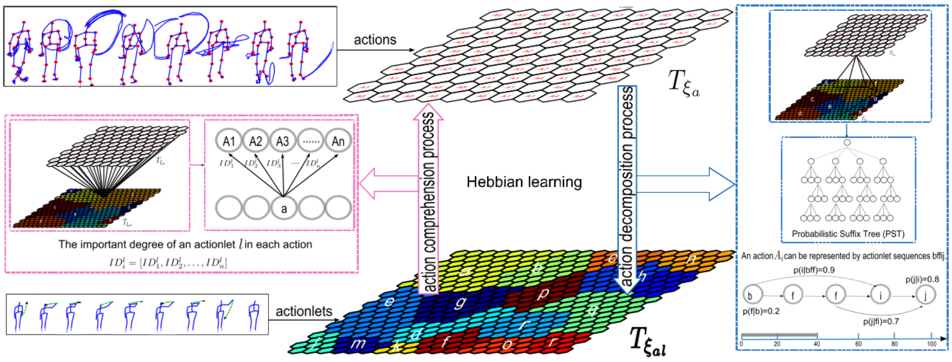

actions
T_{ξ_a}
A1
A2
A3
......
An
a
The important degree of an actionlet l in each action
action comprehension process
Hebbian learning
action decomposition process
Probabilistic Suffix Tree (PST)
An action A_i can be represented by actionlet sequences bffij
p(i|bff)=0.9
p(j|i)=0.8
p(f|b)=0.2
p(j|fi)=0.7
actionlets
$T_{\xi_{al}}$

Figure 2.25 – *Overview of the method of [34]*

tation by comparing the similarity between human skeletal joint trajectories in a Riemannian manifold [55] and a comprehensive survey [44] of existing space-time representations of people based on 3D skeletal data.

2.5.4 Methods based on hybrid information

Here we introduce some works proposed hybrid approaches by combining both depth information and skeleton data features in order to improve recognition performances. [59] defined a Markov Random Field MRF over the spatio-temporal sequence where nodes represent objects and sub-activities, and edges represent the relationships between object affordance, their relations with sub-activities, and their evolution over time. This method needs the video to be pre-segmented. And the object detection is independent of the contextual feedback from human actions.

[117] used relative skeleton position and local occupancy patterns (LOP) features to model the human-object interaction, and developed Fourier Temporal Pyramid to characterize temporal dynamics. [120] proposed a 4D human-object interaction model (4DHOI) for event recognition

and object detection. They model the 4D human-object interaction with a hierarchical graph, as Figure. 2.26 shows. In addition, they collected a new multi-view dataset named Multiview 3D Event Dataset with 8 types of human object interaction.

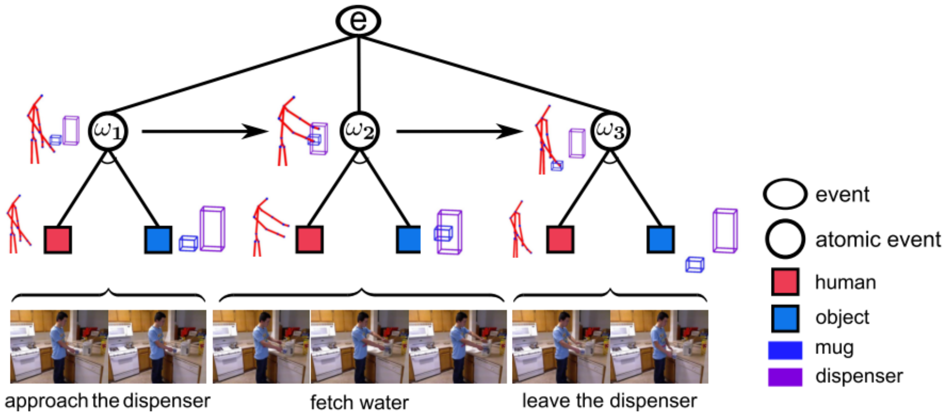


Figure 2.26 – *Hierarchical graph model of event[120]*

In the work of [97], they adopted a combination of multi-modal multi-part features to recognize human activities as shown in Figure. 2.27. The multi-modal multi-part features learned by the proposed hierarchical mixed norm which used for a group feature selection.

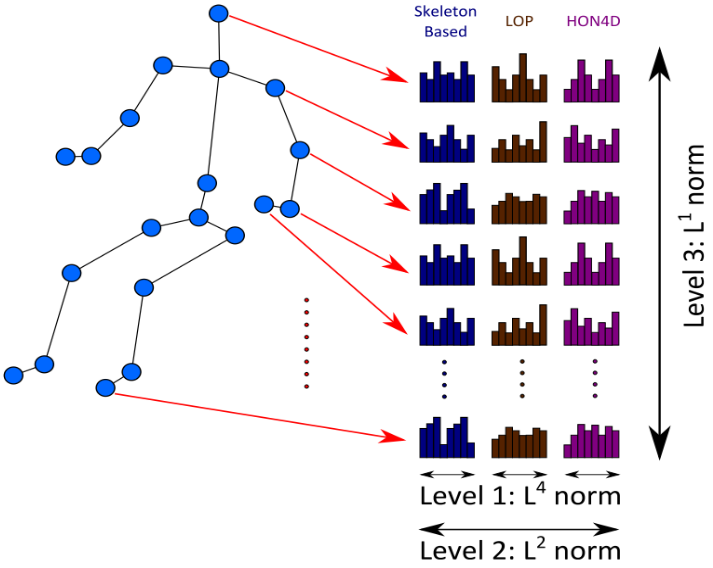


Figure 2.27 – *Three Levels of the Proposed Hierarchical Mixed Norm for Multimodal Multipart Learning. [97]*

Different from [97], [110] utilized more depth information. They used spatio-temporal features of segmented body depth map and body joint features for activity recognition learning by Hidden Markov Models (HMM). [125] focused on daily activity recognition which using the feature extracted from skeletal and depth data. [141] adopted color and depth data to build their feature named CoDe4D LST features that are extracted in 4D space.

All of the methods of [97] [110] [125] [141] were performed on MSR-Daily Activity dataset which we adopt in our framework.

2.6 Related work concerning abnormal gait detection

Now human abnormal gait detection attracts more concern for earlier detection of human diseases. For example, falls on the stairs are a common cause of accidental injury especially among the older adults. So to understand the mechanisms behind such accidents is significant for the prevention of falls, and the support of independent living among elderly. In the sense, the present research aims to apply the recent improvements in human gait analysis based on low-cost RGB-D devices.

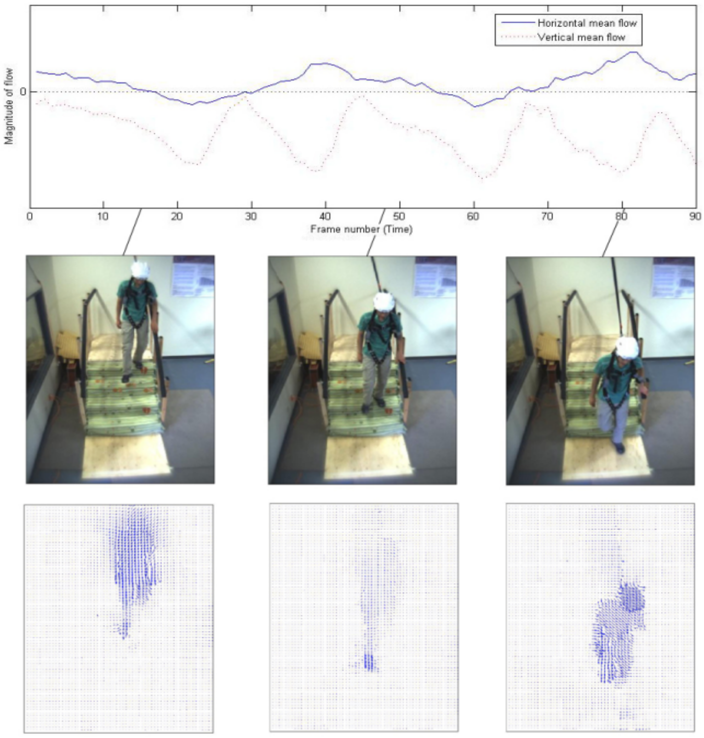


Figure 2.28 – *Optical flow computation for an example stair descent. There are the mean vertical and horizontal optical flow, original images, and resulting optical flow in the three rows. [100]*

In the work of [100], they used monocular RGB images to detect unusual human events on stairs. They tracked both feet by using a mixed state particle filter, and computed two different sets of features to classify stairs descents using a hidden Markov model: the foot positions and velocities and the parameters of the mean optical flow over a foreground region.

For the same issue, [86] analyzed the abnormal gait from skeletons information. They used binary classifiers of harmonic features to detect abnormalities in stairs descents from the skeleton joints of Kinect. The 3D skeleton joints, provided by Kinect, were adopted to estimate the walking speed and extract the feature by encoding human motion during stairway descent. They implemented fall detection automatically. Compared to previous research which utilized feet identified visual tracking as the best feature of dangerous activities, 3D motion of the hips was proved by experimental results that was the most relevant feature in detecting abnormal gait experimentally shown to be the most informative component in detecting abnormal gait.

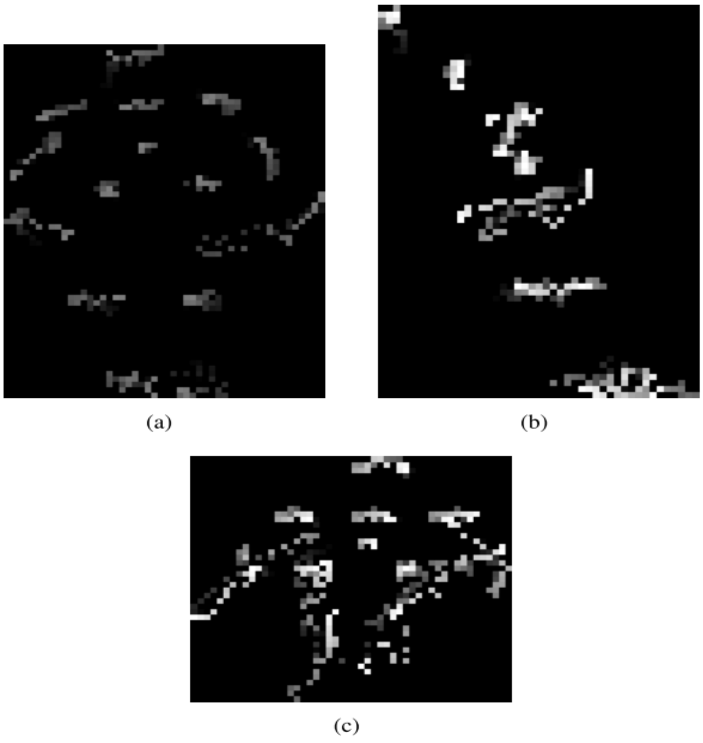


Figure 2.29 – *Sample of front (a), side (b) and bottom (c) views that make up a specific instance of the JMH feature. [21]*

Along similar lines, [21] detected joint motion history feature based on RGB-D devices and used the BagOfKeyPoses to classify temporal feature sequences JMH (Figure. 2.29) for abnormal gait detection. They recorded their own dataset for abnormal gait detection by Microsoft Kinect V2 which obtained 3D skeleton data that was only input data for spatio-temporal features. The spatio-temporal features were learned by a bag of key poses model. Then reduced the dimensionality by axis projection to get JMH feature which be classified to detect abnormal behaviors. Specially, the protocol of this work is different from other works: their training set only included normal class but testing set included abnormal and normal samples. In our work, We use the dataset to test our approach.

2.7 Conclusion

As discussed above, the Kinect sensor changes the picture of human activity analysis, especially brings many advantages for human object interaction recognition. From recent research works, many researcher turned to the study of depth or skeleton information rather than RGB data for human object interaction or abnormal gait analysis.

Depth information based approaches implement the description of actions and interactions by using their geometrical data which is global or local. Note that some approaches adopted global sequence which is informative to analyze the whole video of human activities. Indeed, this kind of approaches show their advantage of avoiding occlusions and noises. However, they are less efficient than the approached that adopted local depth data of human activities. Specially, the real time issue emerges in recent research.

Skeleton information estimated from depth map brings benefits to human object interaction and abnormal gait recognition. Modeling the skeleton information facilitates online detection for these tasks. Still, there are limitations of skeleton data in the cases of human object interaction. The object details can not be described well when recognizing similar interactions like using laptop and reading book. So the depth information can hand it well for complex behaviors. Due to this issue, hybrid information combining strengths of skeleton and depth data are appreciated, where the real time recognition hardly to implement.

The balance between skeleton and depth to achieve efficient and accurate human object interaction recognition is the challenges we should be concerned about. Hence, a better solution should be explored for human object interaction understanding.

Thus, in the following chapter, we investigate this issue and propose a spatio-temporal modeling of human object interaction videos for on-line and off-line recognition. Meanwhile, the on-line frame work is applied on solving the problem of detecting abnormal gait.

First, spatial modeling framework is adopted for generating two kinds of features: low-level feature and object feature. Second, temporal modeling frame is used for solving the issue about rate-invariant. Finally, the sequences are modeled for two scenarios that are on-line classification and rate-invariant classification for off-line. Both scenarios adopt Random Forest classifier to implement human object interaction recognition.

3 Spatial Modeling

Sommaire

3.1 Introduction

Generally, an action or activity recognition system can be viewed as proceeding from a sequence of images to a higher-level interpretation in a series of steps [109]. The major steps involved are the following:

1. Input video or sequence of images
2. Extraction of concise low-level features (e.g. tracking and object detection)
3. Mid-level action descriptions from low-level features (e.g. activity recognition modules)
4. High-level semantic interpretations from primitive actions.

In the literature of activity recognition, most of the previous works have focused on simple human action recognition such as boxing, kicking, walking, etc. However, human activity understanding is a more challenging problem due to the diversity and complexity of human behaviors [138] and accurate human action recognition is still a quite challenging task and is gradually moving towards more structured interpretation of complex human activities involving multiple people and especially interaction with objects. Actually, during a human object interaction scene, the hands may hold objects and are hardly detected or recognized due to heavy occlusions and appearance variations [119]. A high level of information of the objects is needed to recognize the human-object interaction.

In this chapter, we introduce three kinds of feature to represent human object interaction. First, we adopt pairwise distance between skeleton joints as low-level feature to represent human action. Then, we use LOP algorithm to detect object location and propose object related distance feature. Last, for a better description of human object interaction, we modify LOP algorithm to obtain the more information of object and further to achieve object feature.

3.2 Low-level feature

The invariance to the translation and rotation of the subject in the scene is a necessary condition of human-object interaction recognition systems: two instances of the same action differing only for the position and orientation of the person with respect to the scanning device should be recognized as belonging to the same action class.

This goal can be achieved either by adopting a translation and rotation invariant representation of the action sequence or providing a suitable distance measure that copes with translation and rotation variations.

We propose to use the inter-joints distances that handles well with the situations discussed above. The skeleton information is denoted as J which contains n joints from the original skeleton data.

$$J = \{j_1, j_2, ..., j_n\} \tag{3.1}$$

d refers to the set of the pairwise distances between the joint a and joint b from J.

$$d = \{d(a,b)\}, a \in J, b \in J \tag{3.2}$$

In Fig. 3.1, there are examples of pairwise distances between skeleton joints of two human object interactions: one is using remote and the other is reading book. The red lines represent the pairwise distances between each joints.

3.3 Object related distance feature

For completely describing human object interaction, the object detection is very important step during the whole feature extraction procedure. Here, we adopt Local Occupancy Patterns (LOP) [117] to find the object location. Then, we obtain our object related distance feature by utilizing the position information of object.

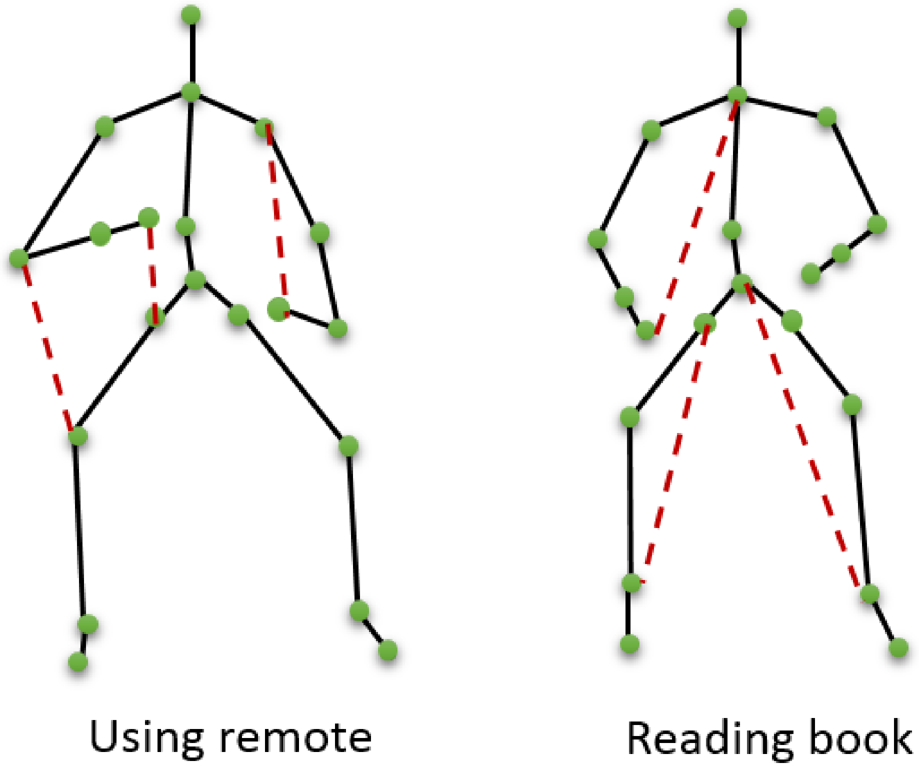


Figure 3.1 – *Examples of pairwise distances between skeleton joints*

3.3.1 LOP algorithm

The Local Occupancy Patterns feature obtained by utilizing the local interaction information that is local depth information around a particular joint. Hence, the interaction can be characterized by this type of information.

In each frame t, there is 3D point cloud generated from the depth map of each frame. The local space of each joint j is represented by $N_x \times N_y \times N_z$ spatial gird. The grid size for each bin is $(S_x \times S_y \times S_z)$ pixels. So the point of each bin b_{xyz} is counted for obtaining the feature o_{xyz} of it. There is a sigmoid normalization function applied. As a result, the local occupancy data of this bin is

$$o_{xyz} = \delta\left(\sum_{q \epsilon b_{xyz}} I_q\right) \tag{3.3}$$

$I_q = 1$ refers to a point in the location q . $I_q = 0$ refers to no point in this location. $\delta(.)$ is the function which is sigmoid normalization:

$$\delta(x) = \frac{1}{1 + e^{-\beta x}} \tag{3.4}$$

The feature vector o_{xyz} of the whole bins in the spatial gird around this joint i is the LOP feature which denoted by o_i.

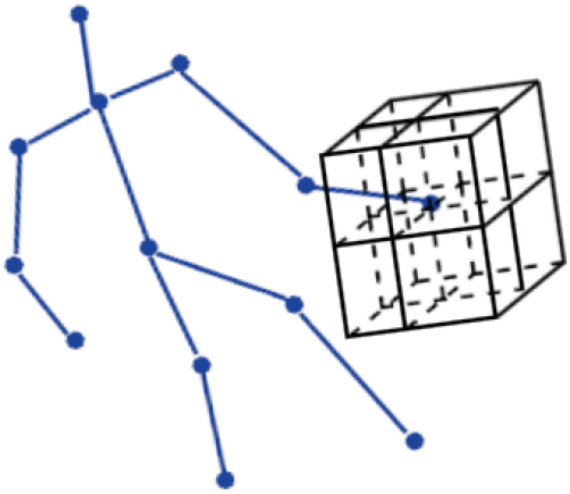


Figure 3.2 – *Illustration of Local Occupancy Patterns (LOP) algorithm*

3.3.2 Object representation

The object position is detected by the LOP algorithm. For each frame, all pairwise distances of 20 skeleton joints and object one are calculated. When the action does not have object, the corresponding entries in the distance matrix are blank and are filled using an imputation technique [10].

In our framework, we employed the mean imputation method, which consists of replacing the missing values by the means of values already calculated in presence of the object from the training set. The skeleton and object information is denoted as J_o which contains 20 joints from the original data and object joint represented by j_o.

$$J_o = \{j_1, j_2, ..., j_{20}, j_o\} \tag{3.5}$$

Here, D_o refers to the set of the pairwise distances between the joint a

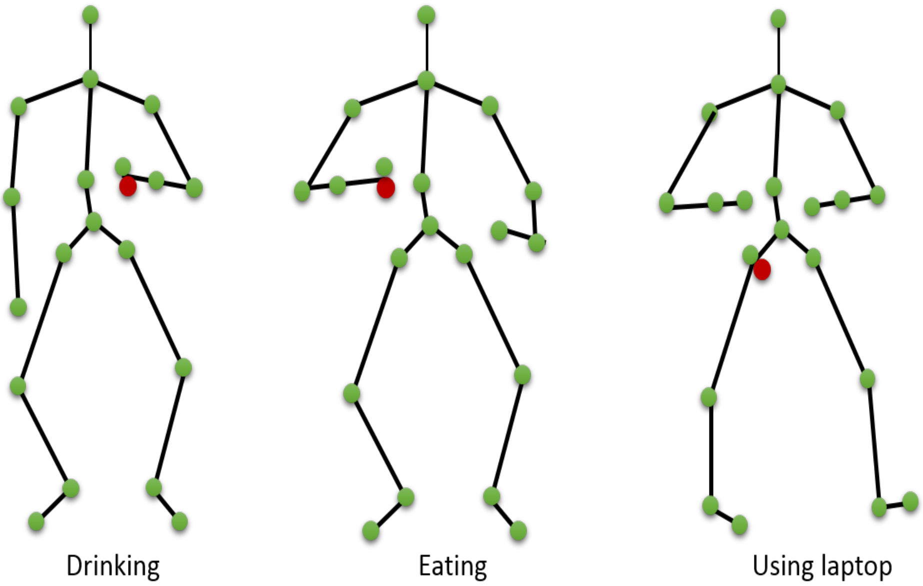


Figure 3.3 – *Examples of our pairwise joint distance features on MSRDaily Activity 3D Dataset. The red one refers to object joint for each action.*

and joint b from J_o.

$$D_o = \{d(a,b)\}\,, a \in J_o\;, b \in J_o \tag{3.6}$$

Thus the low-level feature vector is composed by the all pairwise distances between the joints and the distances between the object and the joints. The size of this vector is equal to $m \times (m-1)/2$, with $m = 21$: the 20 joints and the object joint. The concatenation of this feature vector along frames gives rise to a trajectory. The shape of the resulting trajectories will be investigated in the later sections for human object interaction classification.

We report in Fig. 3.3 examples of different human action interactions. The object is reported in red and the joints are reported in green. The proposed features are the pairwise distances between all these points in 3D.

3.4 Object feature

It is insufficient to only use the 3D joint positions to fully model an action, especially when the action includes the interactions between the subject and other objects. Therefore, it is necessary to design a feature to describe the local depth appearance for the joints.

For tracking of hands in action or in interaction with objects provides rich information that can be used to interpret a number of human activities. In this section, we propose an novel feature based on depth information around two hands for more complicated human object interaction analysis.

Semantically, human-object interactions are relevant to characters of objects for each of them. That is why many works devoted themselves to object detection for recognizing interactions. But this is a difficult and time consuming way to realize online classification. As we discussed in the previous part, it is insufficient to only use the 3D joint positions to fully model an action, especially when the action includes the interactions between the subject and other objects such as *drinking* and *picking phone*. The extra input like depth information need be adopted in order to have more precise classification.

Motivated by properties of objects, we try to utilize the size and shape information of objects which is more efficient and convenient way for online human-object interaction recognition. When performing an interaction, human usually hold objects by two hands. Moreover, the depth points located around the skeleton joints of two hands contain a lot of messages about the size and shape of objects.

3.4.1 Modified LOP

Different from LOP algorithm, we modified the LOP algorithm using depth information around two hands for efficient human object interaction modeling.

As with LOP algorithm, there is 3D point cloud generated from the depth map of each frame t. We will utilize the information from the local space of each hand joint j is represented by $N_x \times N_y \times N_z$ spatial gird. The grid size for each bin is $(S_x \times S_y \times S_z)$ pixels.

In the next part, we will discuss about the optimization of the grid size.

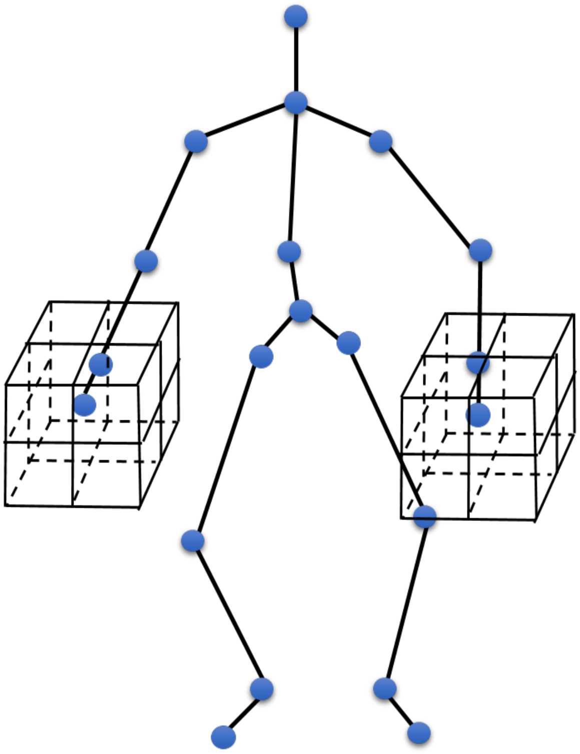


Figure 3.4 – *Illustration of modified LOP algorithm*

3.4.2 Size estimation

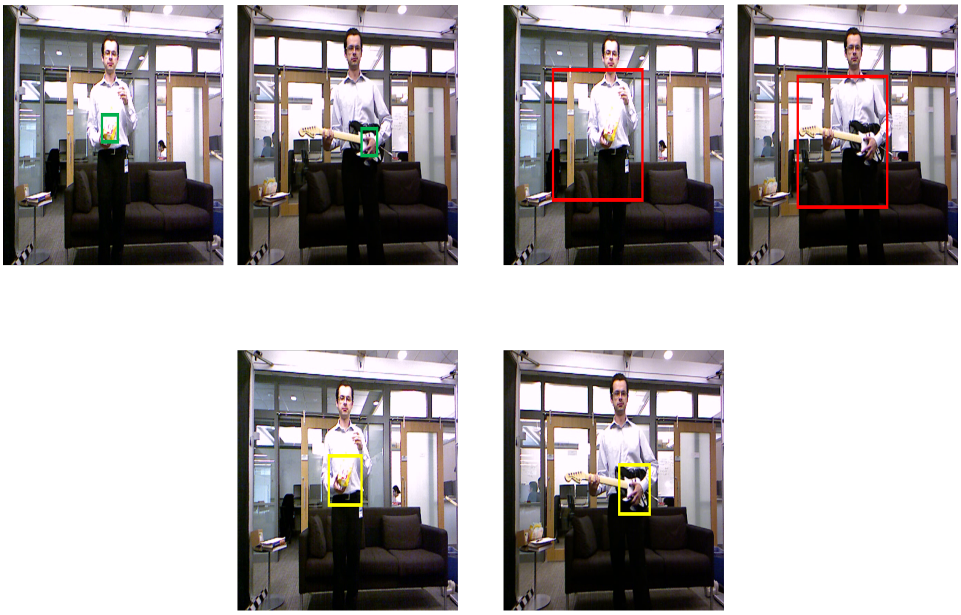


Figure 3.5 – *The illustration of object cube size. In the first row, the green rectangular and red rectangular represent small and large cube respectively. The yellow rectangular in the second row represent the appropriate size of object cube.*

The feature vector calculation depends on the size of chosen cubes to detect these points. If the cube size is too small like the situation shown at the top left in Fig. 3.5, the green rectangular is too small to show the features of different object. So the resulting feature will not be discriminative for interaction classification. If the cube size is too big like the situation shown at the top right in Fig. 3.5, the red rectangular is so big that contains a lot of context from background and other parts of body. So we have to detect object in a appropriate size as shown in the second row in Fig. 3.5. In the experiment, the retained size of cubes is 50. A trade of the size of this cube and the results will be discussed later in experimental section.

3.4.3 Shape estimation

For estimating the rough shape of each object in the human object interactions, we apply principal component analysis (PCA) on object features. Before discussing the detail of rough shape about objects, we briefly introduce the principle of PCA.

PCA

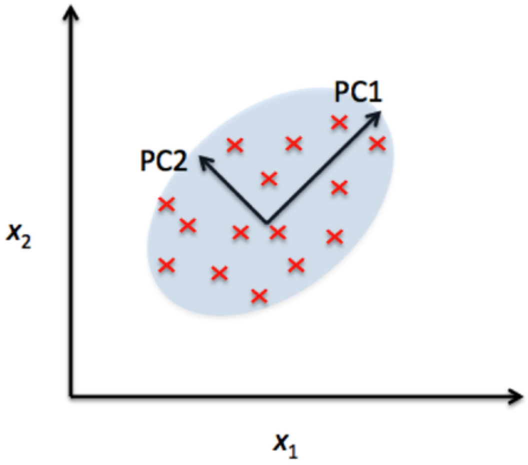


Figure 3.6 – *Illustration of principal component analysis*

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less

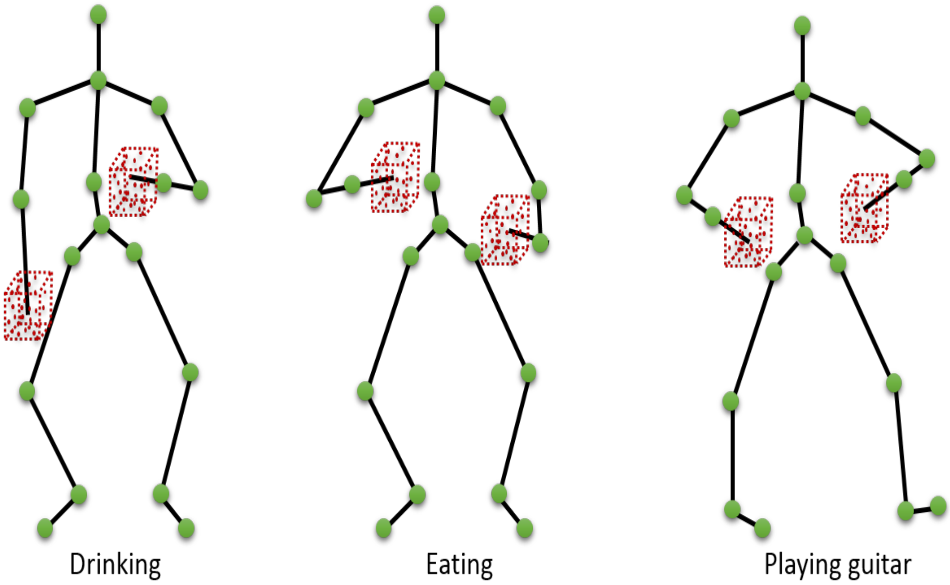


Figure 3.7 – *Examples of our object features on 3D Online Action dataset. The red cube refers to object cube for each action.*

than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

Shape estimation

The object is assumed to be present around one hand, thus similarly to the LOP algorithm [117] that counts the number of points inside a given cube around given point (hand for example) and decides the presence of an object given a threshold, we extend this algorithm to exploit the number of the points inside the cube and the 3D coordinates of these points to built the object feature.

The number of depth points refers to the rough size of objects and the coordinates of these points refer to the rough shape of objects. The PCA algorithm is applied on the coordinates in order to determine the principal directions of the object inside the cube. These directions are concatenated with the number of the points to built the object feature.

3.5 Conclusion

In this chapter, we have brought in the conceptions about low-level feature and objection feature. The low-level feature consists of pairwise distance between human skeleton joints and object related distance feature that uses LOP algorithm to detect object location. Except the feature only based on skeleton data, we propose object feature based on describing the rough shape and size of objects which include depth information around two hands. This proposed object feature can be useful for better description the relationship between human pose and object.

Due to the different properties of these two features, they are adopted in the different applications which will be introduced in further detail.

Temporal Modeling 4

Sommaire

4.1 Introduction

As introduced in the beginning of the last chapter, there are several steps for human object interaction recognition system which can be viewed as proceeding from a sequence of images to a higher-level interpretation. In the chapter, we will bring in the higher level interaction description to meet the different challenge: rate invariance.

We start by outlining a mathematical framework for helping in analyzing the temporal evolution of human object interactions when viewed as trajectories on shape space of distance trajectories. This framework respects the underlying geometry of the shape space of the trajectories and helps maintain desired invariance. Then calculate the distance between the mean trajectory of each action to the trajectories from testing set based on square-root velocity function (SRVF) [102].

So that the trajectories from different action classes can be fairly compared in a another shape space.

4.2 Distances Evolution Modeling and problem position

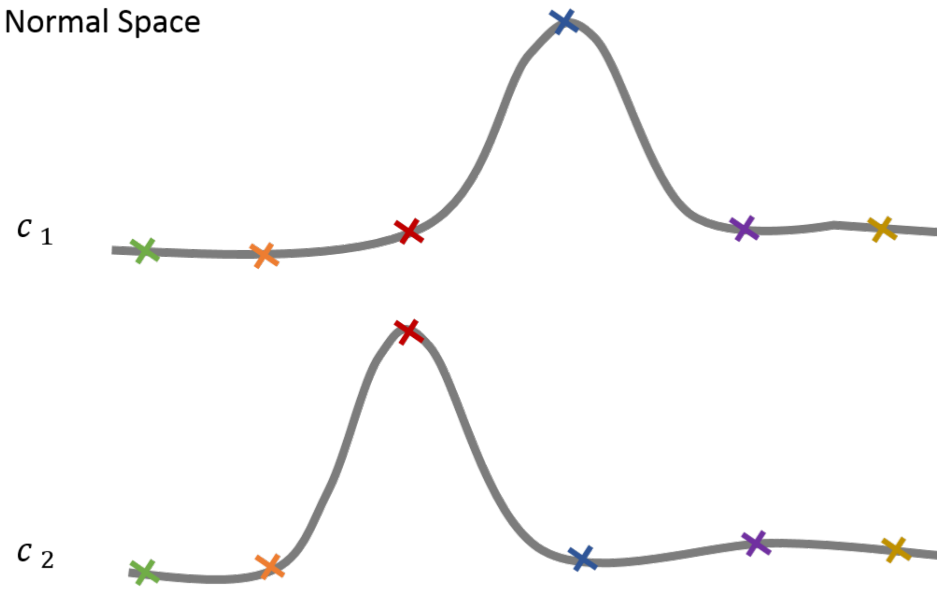


Figure 4.1 – *An illustration of sequences alignment in normal space*

For human object interaction recognition, one of the challenges is the variations in the execution speed of the action. So a rate invariant frame work for modeling human object interaction temporally is proposed. It is inaccurate if we intend to align two sequences in normal space directly. There is an illustration Fig. 4.1 for explaining this situation. We can see that the locations of these points on curve c_1 which refers to human object interaction are not changing after alignment on normal space. However, the corresponding location of curve c_2 which refers to human object interaction after alignment on normal space are totally different from c_1. So we will not compare the sequences in this space in order to be invariant to the rate of execution.

In general, the problem of sequence alignment is reduced to a problem of high dimension curves reparametrization. For more details, please refer to appendix. So the square-root velocity function (SRVF) is used for such task. After SRVF, the curve c_1 is are represented as q_1 in a higher dimensional space that we call it shape space here. The corresponding location of each point remains the same after alignment.

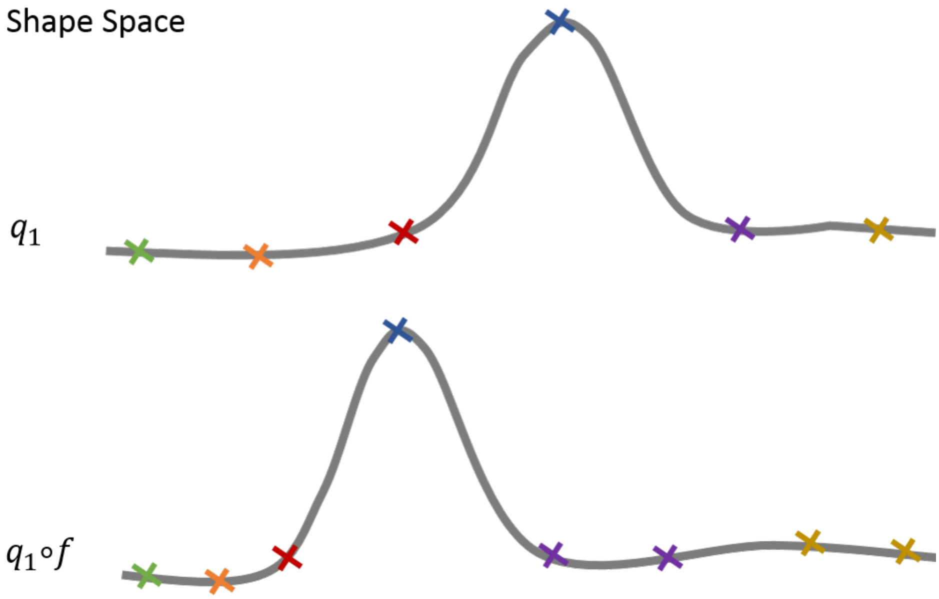


Figure 4.2 – *An illustration of sequences alignment in shape space*

4.3 Background on Shape Analysis Framework

Here, we briefly introduce the duty of square-root velocity function (SRVF) and Karcher mean on a manifold M.

4.3.1 Square Root Velocity Function SRVF

Let $\beta : I \to \mathbb{R}^m$, where $I = [0, 1]$, represents a parameterized curve in $\mathbb{R}^m$.

To analyze the shape of β, we shall represent it mathematically using the *square-root velocity function* (SRVF) [103], denoted by $q(t)$, according to: $q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}$; $q(t)$ is a special function of β that simplifies computations under elastic metric.

Actually, under $\mathbb{L}^2$-metric, the re-parametrization group acts by isometries on the manifold of q functions, which is not the case for the original curve β. To elaborate on the last point, let q be the SRVF of a curve β. Then, the SRVF of a re-parameterized curve $\beta \circ \gamma$ is given by $\sqrt{\dot{\gamma}}(q \circ \gamma)$. Here $\gamma : I \to I$ is a re-parameterization function and let Γ be the set of all such functions.

Define the preshape space of such curves: $\mathcal{C} = \{q : I \to \mathbb{R}^m | \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^m)$, where $\| \cdot \|$ implies the $\mathbb{L}^2$ norm. With the $\mathbb{L}^2$ metric on its tangent spaces, $\mathcal{C}$ becomes a Riemannian manifold. Also, since the elements of $\mathcal{C}$ have a unit $\mathbb{L}^2$ norm, $\mathcal{C}$ is a hypersphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^m)$. The geodesic path between any two points $q_1, q_2 \in \mathcal{C}$ is given by the great circle, $\psi : [0, 1] \to \mathcal{C}$, where

$$\psi(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2\right), \tag{4.1}$$

and the geodesic length is $\theta = d_c(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle)$.

In order to study *shapes* of curves, one identifies all re-parameterizations of a curve as an equivalence class.

Note that the parameterization of a trajectory during an action corresponds to the rate of the action. Thus comparison of equivalent classes rather than trajectories themselves is rate invariant differentiation which reduces the difference in rate between actions and facilitates the action recognition.

Let's define the equivalent class of q as: $[q] = \{\sqrt{\dot{\gamma}(t)}q(\gamma(t)),\ \ \gamma \in \Gamma\}$. The set of such equivalence classes, denoted by $\mathcal{S} \doteq \{[q]|q \in \mathcal{C}\}$ is called the *shape space* of open curves in $\mathbb{R}^m$. As described in [103], $\mathcal{S}$ inherits a Riemannian metric from the larger space $\mathcal{C}$ due to the quotient structure. To obtain geodesics and geodesic distances between elements of $\mathcal{S}$, one needs to solve the optimization problem:

$$\gamma^* = argmin_{\gamma \in \Gamma} d_c(q_1, \sqrt{\dot{\gamma}}(q_2 \circ \gamma)). \tag{4.2}$$

The optimization over Γ is done using the dynamic programming algorithm. Let $q_2^*(t) = \sqrt{\dot{\gamma^*(t)}}q_2(\gamma^*(t)))$ be the optimal element of $[q_2]$, associated with the optimal re-parameterization γ^* of the second trajectory, then the geodesic distance between $[q_1]$ and $[q_2]$ in $\mathcal{S}$ is $d_s([q_1],[q_2]) \doteq d_c(q_1, q_2^*)$ and the geodesic is given by Eqn. 4.1, with q_2 replaced by q_2^*.

4.3.2 Karcher mean

One advantage of a shape analysis framework of the trajectories is that one has the actual deformations in addition to distances. In particular, we have a geodesic path in $\mathcal{S}$ between the two trajectories β^1 and β^2 in $\mathbb{R}^m$. This geodesic corresponds to the optimal elastic deformations of two trajectories. The Riemannian structure defined on the manifold of shape of the trajectories in $\mathcal{S}$ enables us to perform such statistical analysis for computing curves (trajectories) mean and variance. The Karcher mean utilizes the intrinsic geometry of the manifold to define and compute a mean on that manifold. It is defined as follows: Let $d_s(\beta^i, \beta^j)$ denote the length of the geodesic from β^i to β^j in $\mathcal{S}$.

To calculate the Karcher mean of trajectories $\{\beta^1, ..., \beta^n\}$ in $\mathcal{S}$, define the variance function:

$$\mathcal{V} : \mathcal{S} \to \mathbb{R}, \mathcal{V}(N) = \sum_{i=1}^{n} d_s(SRVF(\beta^i), SRVF(\beta^j))^2 \tag{4.3}$$

The Karcher mean is then defined by:

$$\overline{\beta} = \arg\min_{\mu \in \mathcal{S}} \mathcal{V}(\mu) \tag{4.4}$$

The intrinsic mean may not be unique, i.e. there may be a set of points in $\mathcal{S}$ for which the minimizer of $\mathcal{V}$ is obtained. To interpret geometrically, $\overline{\beta}$ is an element of $\mathcal{S}$, that has the smallest total deformation from all given trajectories.

Algorithm 1: Karcher mean algorithm

```
Set k = 0. Choose some time increment $\epsilon \leq \frac{1}{n}$. Choose a point $\mu_0 \in \mathcal{S}$ as an initial
  guess of the mean. (For example, one could just take $\mu_0 = \beta^1$.)

    1- For each $i = 1, ..., n$ choose the tangent vector $t_i \in T_{\mu_k}(\mathcal{S})$ which is tangent to the
    geodesic from $\mu_k$ to $\beta^i$. The vector $g = \sum_{i=1}^{i=n} t_i$ is proportional to the gradient at $\mu_k$ of
    the function $\mathcal{V}$.
    2- Flow for time $\epsilon$ along the geodesic which starts at $\mu_k$ and has velocity vector $g$. Call
    the point where you end up $\mu_{k+1}$.
    3- Set $k = k + 1$ and go to step 1.
```

4.4 Trajectories classification

Note that the both training and testing data are built by spatio-temporal modeling and the red point is the object position we assumed. First, Spatio-Temporal Modeling (STM) is applied on each video of training and testing data to get trajectories of dimension $\mathbb{R}^{m*n}$ (where n is the number of frames for each video). Then, the rate-invariant mean shape μ_i of each action $a_i, i = 1..k$ is calculated. The feature vector for a given trajectory is then built by concatenating the distances d_S between this trajectory and all of the mean trajectories.

4.4.1 Pre-processing: Trajectories re-sampling and smoothing

Since the trajectories contains a number of imperfections, such as spikes. The data pre-processing is very important and non-trivial. We use smoothing filter which reduces high frequency components (spikes) in the trajectories, improves the shapes of trajectories. Here, it is the resampling formula we used in this work:

$$\beta(t) = 0.25\beta(t-1) + 0.5\beta(t) + 0.25\beta(t+1) \tag{4.5}$$

We also did the test on the curves with resampling or not, and the curves with smoothing or not. Table. 4.1 shows the performance of the 3D online action [139] when we utilized different pre-processing methods.

Table 4.1 – *The performance of different pre-processing methods*

Pre-processing	Recognition Rate
Only after resampling	72%
Only after smoothing	75.5%
After resampling and smoothing	76%

4.4.2 Sequence Classification

We will apply the framework we introduced above on our low level feature extracted from skeleton data of each frame which dimension is $\mathbb{R}^{210}$.

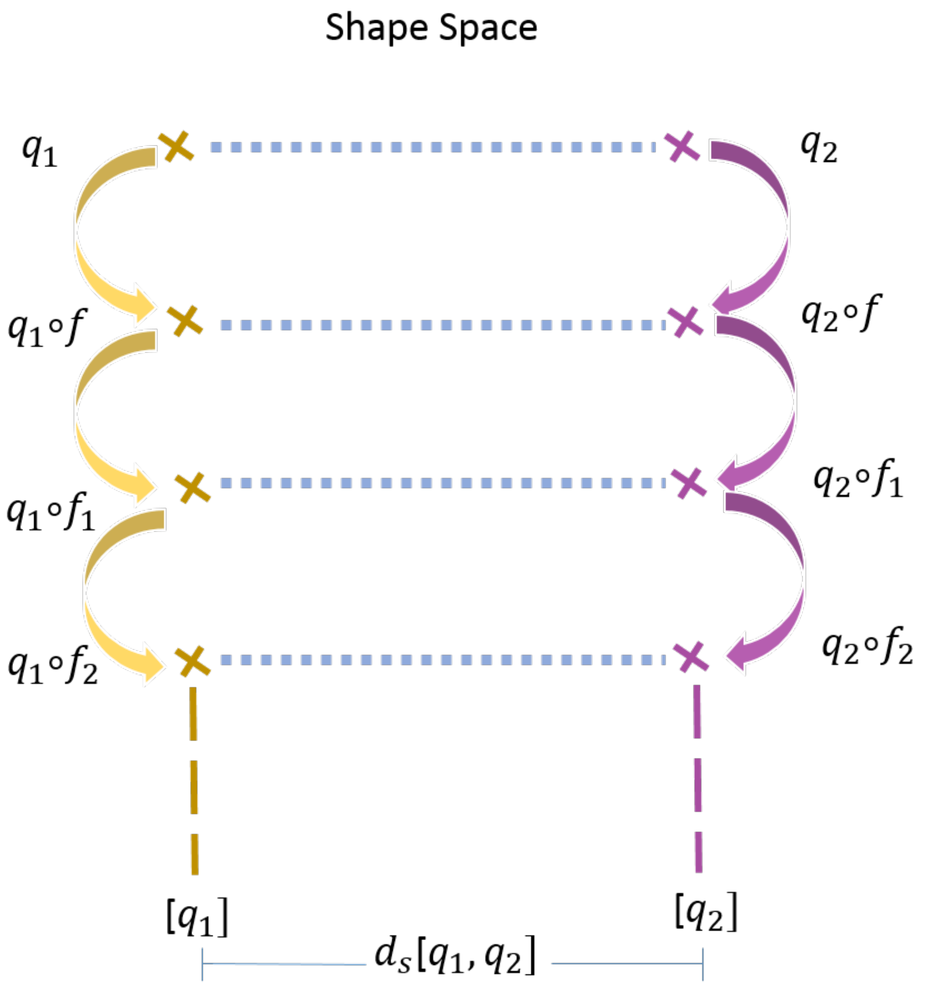


Figure 4.3 – *An illustration of sequences alignment in shape space*

In this way, the distance between the shape of two curves in $\mathbb{R}^m$ is invariant to their translation, scale, rotation and re-parametrization. f is reparametrization function.

4.4.3 Mean shape of trajectories and feature vector calculation

The mean are calculated on trajectories belonging to the same action in order to get mean of the trajectory for each action. These means will be used in the classification of the trajectories. Moreover, the mean trajectory is invariant to the rate of execution of given videos due to the elastic metric used in the calculation of the mean.

The main step of feature vector calculation is shown in Fig. 4.4. The feature vector is built by using the distances to the means of the actions calculated on train data. Given train set $T = \{\beta^1, ..., \beta^n\} \in \mathbb{R}^{210*n}$, each trajectory corresponds to an action class $label_i \in \{a_1, ..., a_k\}$. We first calculate, using algorithm 1, the mean μ_i for each class. Next, we calculate the geodesic distance d_S between a given curve β and the mean curves. Thus a vector of distance of size k is provided as feature vector to classify the curve β. For example, this is a feature vector size k of one video sequence:

$$d_S = \left\{ d(\beta^1, \mu_1), d(\beta^1, \mu_2), ..., d(\beta^1, \mu_k) \right\} \tag{4.6}$$

4.5 Conclusion

In this chapter, we have introduced the temporal modeling framework that we employs for shape analysis of human object interaction. Within this framework, the shape of a curve is captured using a single representation called the square-root velocity function and interpreted in a manifold called shape space.

In order to compare shapes on such space, we exploit an elastic distance representing similarity between shapes independently to their size, location, orientation and elasticity. As demonstrated in the following, this established temporal modeling framework is the core of our study on human behavior understanding.

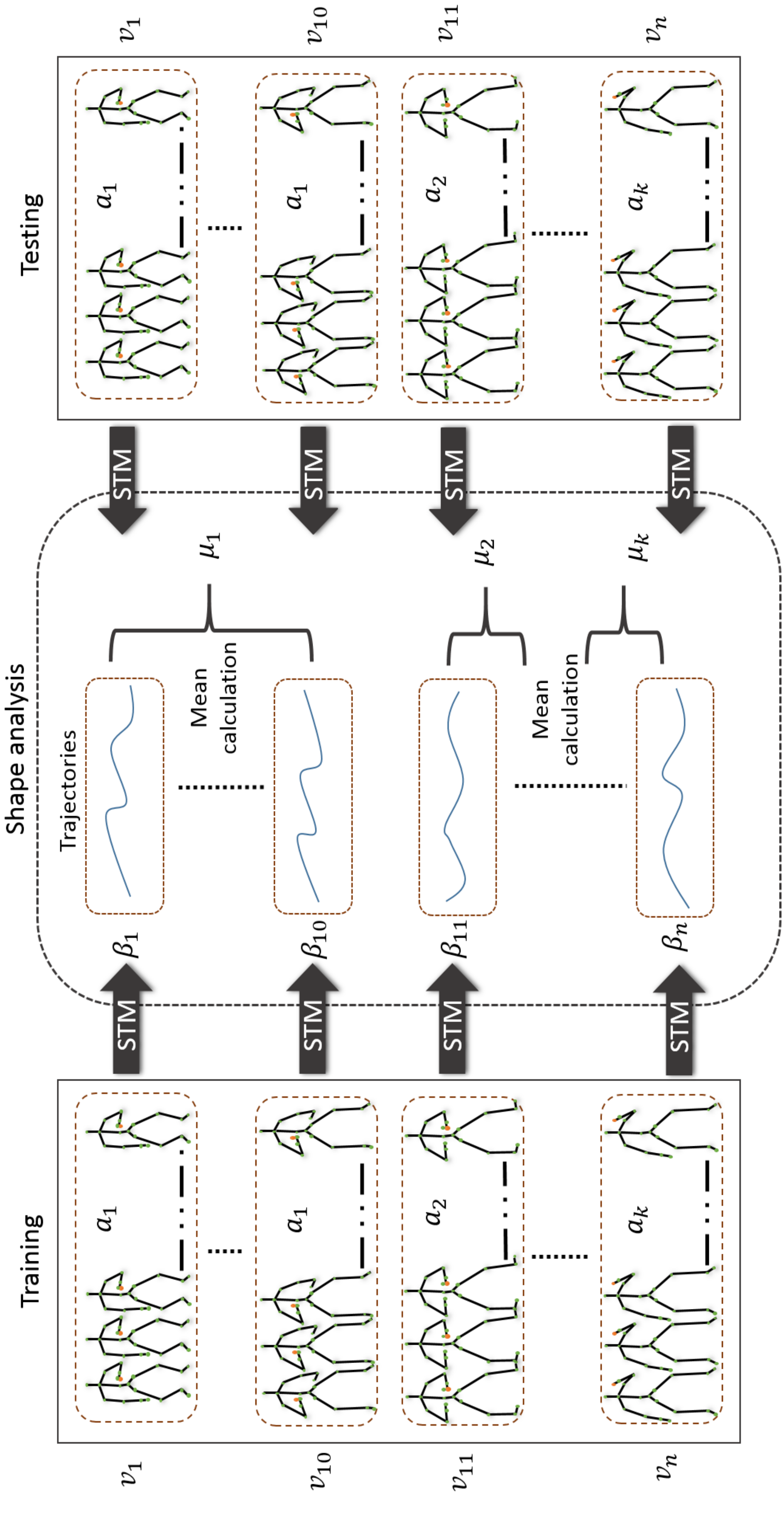


Figure 4.4 – *Overview of off-line classification.*

Human Object Interaction Recognition

5

Sommaire

5.1 Introduction

The goal of building human representations is to extract compact, descriptive information, like features, to encode and characterize a human's attributes from perception data such as gesture, action, and interaction, when developing recognition or other human-centered reasoning systems [44].

In this section, we develop an approach for two different scenarios with different challenges. As show in Fig. 5.1, there are online classification and full sequence based rate invariant classification. The low-level feature adopted on both scenarios due to its property. We will introduce and discuss these two applications and corresponding results in the following parts.

As we can see from Fig. 5.1, there are two main applications: online and off-line classification which are represented by orange and green lines respectively. Here, the two orange lines refer to different frameworks. First, we adopt low-level feature and object related distance feature that already introduced in previous chapter to build our feature vector for classifying task by Random Forest algorithm. Second, we adopt the low-level feature and object feature to build our feature vector by PCA algorithm for such task using the same classifier. The green line referring to off-line classification adopts the low-level feature and object related distance feature which are the input for shape analysis framework to build feature vector for Random Forest-based human object interaction recognition.

Random Forest-based human object interaction recognition

For the classification task we used the multi-class version of Random Forest algorithm. The Random Forest algorithm was proposed by Leo Breiman in [16] and defined as a meta-learner comprised of many individual trees. It was designed to operate quickly over large datasets and more importantly to be diverse by using random samples to build each tree in the forest. Diversity is obtained by randomly choosing attributes at each node of the tree and then using the attribute that provides the highest level of learning. Once trained, Random Forest classify a new action from

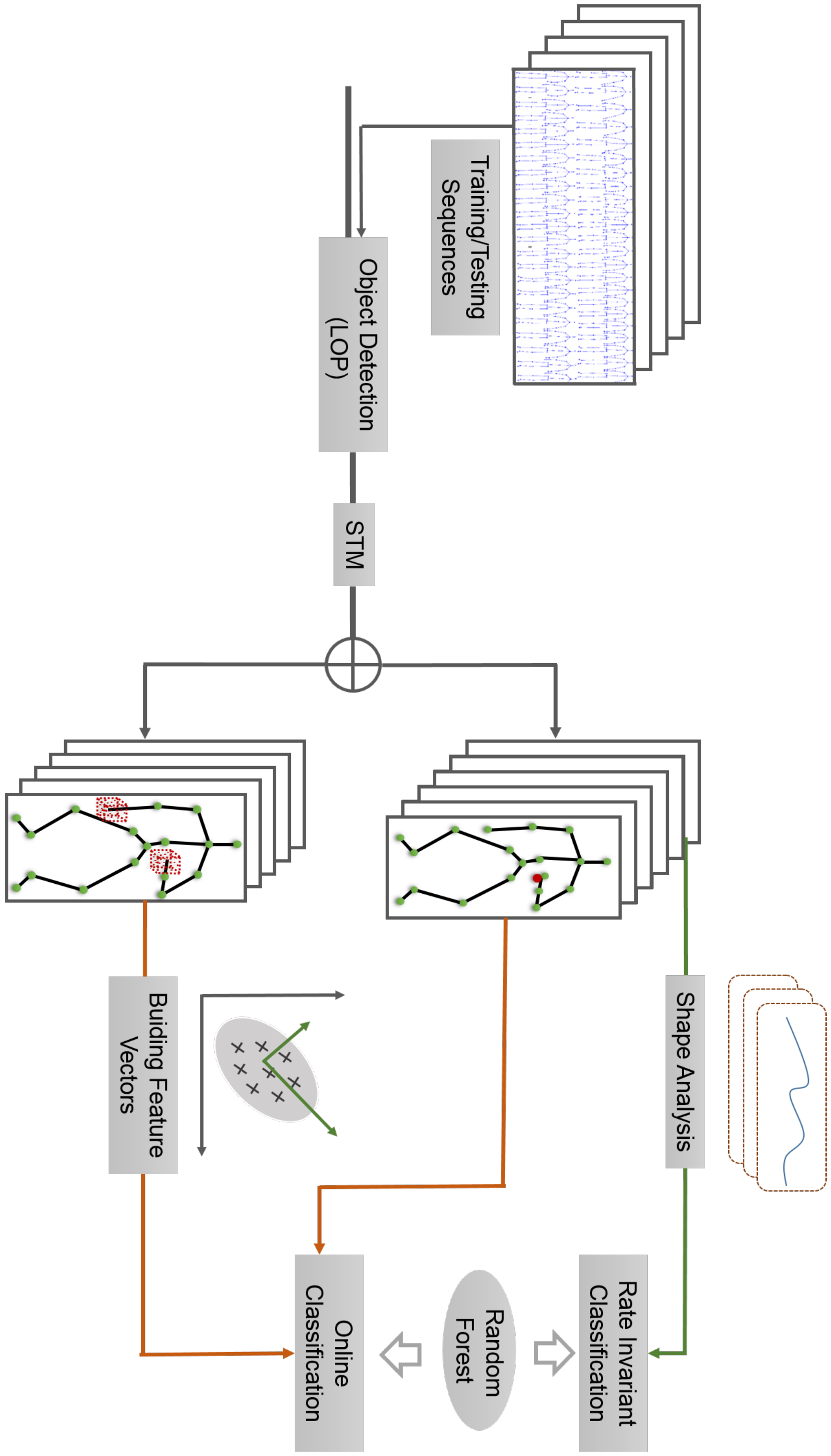


Figure 5.1 – *Overview of our method. The main steps are shown: low-level feature extraction from each frame; Spatial and temporal modeling; Shape analysis of feature vector for rate invariant classification; low-level feature and fused object feature for online classification; Random Forest based classification*

an input feature vector by putting it down each of the trees in the forest. Each tree gives a classification decision by voting for that class. Then, the forest chooses the classification having the most votes (over all the trees in the forest). In our experiments we used Weka multi-class implementation of Random Forest algorithm by considering different trees. A study of the effect of the number of the trees is reported later in the experimental part.

5.2 Online human object interaction classification

Indeed, the trade-off between the accuracy and observation size for rapid and real-time recognition is an important topic in a wide spectrum of real applications, this motivates the first scenario we develop in this thesis: online human-object interaction recognition. The main challenge here is the accurate and real time recognition thus we propose to use low-level features as input to classifier. Actually, the skeleton features (low-level) are easy to extract and track from depth maps thanks to the work of [98], utilizing low level features to describe interactions and the most relevant parts of human poses with respect to object can make it possible to achieve rapid and online recognition of human-object interactions.

The LOP algorithm [117] is applied on each frame of input sequence to detect the presence and the position of the object. Then the low-level features are calculated in each frame and the sequence is modeled as evolution of the resulting feature vector; this step is denoted spatio-temporal modeling. The rough shape and the size of the object detection represent the next step in the pipeline of online human-object interaction recognition. Together with the low-level feature resulting on spatio-temporal modeling, the object feature represent the input of the random forest classifier.

5.2.1 Feature vector building

Using the Spatio-Temporal Modeling (STM) of low-level feature described in the previous chapters, we now develop an online framework with different kinds of object features for helping in analyzing human-object interactions when viewed as trajectories. These object features help maintain desired invariance.

The first step in the on-line recognition system is the object feature extraction. Here, we perform our online framework on two types of feature vectors respectively. First, we only use low-level feature and object related distance feature to build final feature vector for online classification. Second, the object feature will be fused later with the low-level extracted features to built the final features vector which will be classified online. Both of them utilize Random Forest algorithm as classifier.

Dynamic shape deformation analysis

To capture the dynamic of object and skeleton deformations across sequences, we consider the two type of feature vectors computed at n successive frames. In order to make possible to come to the recognition system at any time and make the recognition process possible from any frame of a given video, we consider sub-sequences of n frames as sliding window across the video.

Thus, we chose the first n frames as the first sub-sequence. Then, we chose n-consecutive frames starting from the second frame as the second sub-sequence. The process is repeated by shifting the starting index of the sequence every one frame till the end of the sequence.

The feature vector for each sub-sequence is built based on the concatenation of individual features of the n frames of the sub-sequence.

Thus, each sub-sequence is represented by a feature vector of size the number of distances for one frame times the size of the window n. For the sliding window of size $n \in [1, L]$ that begins at frame i, the feature vector is:

$$x = [V_i, V_{i+1}, ..., V_{i+n-1}] \, ,$$

with L the length of the sequence.

For $n = 1$, our system is equivalent to recognition frame by frame without any memory of previous frames. If $n = L$, the length of the video, our system will provide only one decision at the end of the video.

The effect of the size of the window on the performance is studied later in experimental part.

5.2.2 Classification

3D Online Action dataset

For the online classification evaluation We will evaluate the performance of our approach for action recognition based on the dataset 3D Online Action dataset [139]. This dataset contains seven types of actions which all those actions are human-object interactions: *drinking, eating, using laptop, picking up phone, reading phone (sending SMS), reading book, and using remote*. The bounding box of the object in each frames is manually labeled. In our approach, we use the object labels to locate our object feature. All of the videos are captured by Kinect. Each action was performed by 16 subjects for two times. We compare our approach with state-of-the-art methods on the cross-subject test setting, where half of the subjects are used as training data and the rest of the subjects are used as test data.

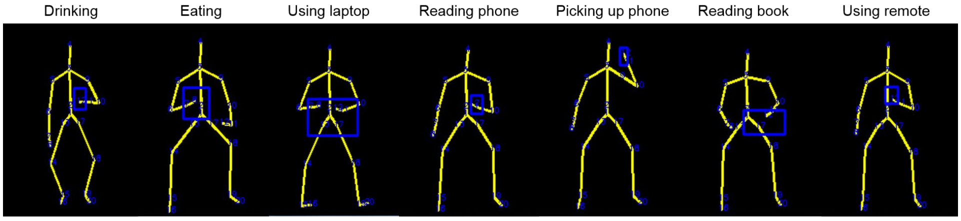


Figure 5.2 – *Skeleton frames from 3D Online Action dataset*

Comparative Evaluation on 3D Online Action dataset

We evaluate the performance of our approach for action recognition based on the dataset 3D Online Action dataset [139]. We compare our approach with the state-of-the-art methods on the cross-subject test setting, where half of the subjects are used as training data and the rest of the subjects are used as test data.

Up to now, our work is the second one based on the 3D Online Action dataset. The first work based on 3D Online action dataset is [139]. The performance presented in this paper uses temporal variation of a joint

Table 5.1 – *Comparison of low-level feature-based online classification on 3D Online Action dataset with state of the art results*

Method	Accuracy
[139] (with memory of previous frames)	71.4%
STM of low-level feature (without memory of previous frames)	72.1%
STM of low-level feature with memory of 10 frames	75.8%

location. We compare our approach based on low-level feature and object related distance feature with the state-of-the-art method on the 2-fold cross-validation and the comparison of the performance is shown in Table 6.1.

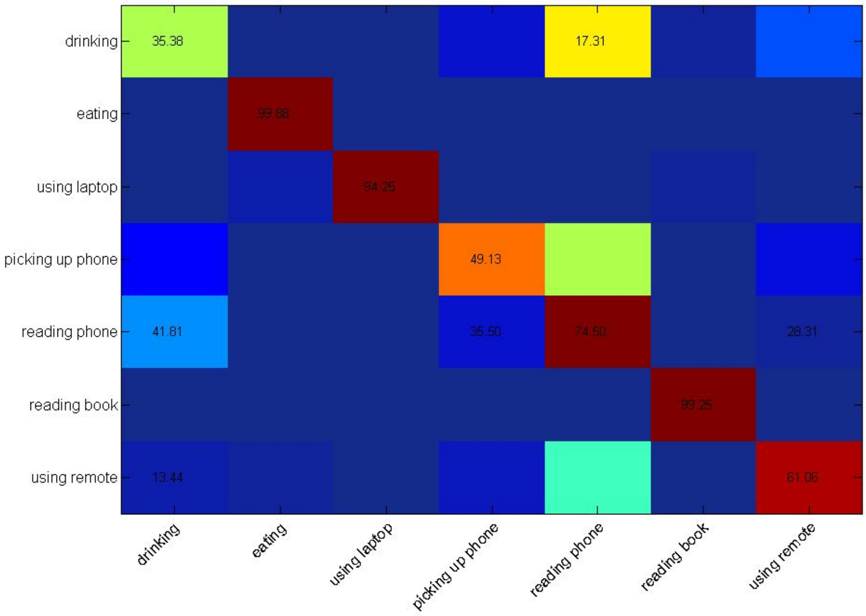


Figure 5.3 – *Confusion matrix using one frame for the proposed approach on 3D Online Action.*

The recognition rate of the discriminative Orderlet Mining is 71.4%, we note that the feature used in their method is with memory of previous frames. Fig. 5.3 shows the confusion matrix without memory of previous frames of the proposed method. Over all seven action categories, 'eating', 'using laptop' and 'reading book' have the best recognition rate. 'drinking' is the most confused action in all cases; it is mostly confused with 'reading phone'.

Actually, they used temporal variation cross frames to build their feature vector. In our method, when $n = 1$, the feature has no memory of previous frames, the mean value of the recognition rate by the Random

Forest is 72.05%. For fair comparison, we need to compare our recognition rate with $n > 1$ to the result in [139]. We achieve 75.8% recognition rate for $n = 10$, which represents better performance than [139]. This result can be due to the use of the distance between the object and the joints that is more relevant than the object position used in [139].

Table 5.2 – *Comparison of object feature-based online classification on 3D Online Action dataset with state of the art results*

Method	Accuracy
[139] (with memory of previous frames)	71.4%
Number of the points in cubes	77.22%
Principle coordinates of the points in cubes	77.26%

As shown in Table 5.2, we compare our methods based on two different object features that are the number of the points and principle direction of these points in specific cube around right and left hands with discriminative orderlet mining [139]. To our best knowledge, this is only work on this database for real time classification. We achieves the classification accuracy based on these two kinds of features are 77.22% and 77.26% respectively. The memory both of them are 50 frames.

Effect of the number of trees in Random Forest algorithm on low-level feature-based classification

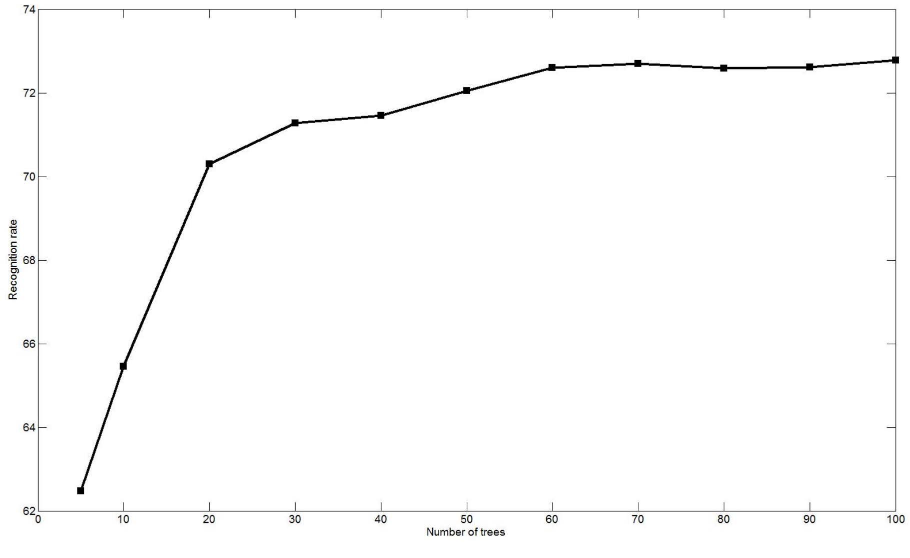


Figure 5.4 – *Human-Object interaction recognition results using a Random Forest classifier when varying the number of trees.*

The performance of Random Forest classifier varies with the number of trees. Thus, we perform the experiments with different numbers of trees; the results of this experimentation is shown in Fig. 5.4. As illustrated in this figure, the recognition rate raises with the increasing number of trees until 60, when the recognition rate reaches 72.5%, and then becomes quite stable. Thus, in the following we consider 50 trees and we report detailed results with this number of trees.

Effect of the temporal size of the sliding window on low-level feature-based classification

We have conducted additional experiments when varying the temporal size of the sliding window used to define the sub-sequences. In Fig. 5.5, we report results for a window size equal to 2, 5 and 10 frames. The recognition rates are respectively 72.8%, 74.8% and 75.8%. Finally, we use the whole length of the sequence (on average this is about 100 frames). From the figure, it clearly emerges that the action recognition rate increases when increasing the temporal length of the window. This reveals the importance of the temporal dynamics and shows that the spatio-temporal analysis outperforms a spatial analysis of the frames. By considering the whole sequences for the classification, the result reaches 82%.

Effect of temporal size of the sliding window on object feature-based online classification

We have conducted additional experiments when varying the temporal size of the sliding window used to define the sub-sequences. We test different sizes of sliding window on the two kinds of feature discussed above and report results on these two datasets for a window size equal to 20, 50 and 80 frames. In Fig. 5.6, there two lines show that: the blue one is the recognition rates of the principle coordinates of the points in cubes on 3D Online Action dataset are respectively 73.74%, 77.22% and 78.27%; the red one is the the recognition rates of the number of the points in cubes on 3D Online Action dataset are respectively 74.11%, 77.26% and 77.81%.

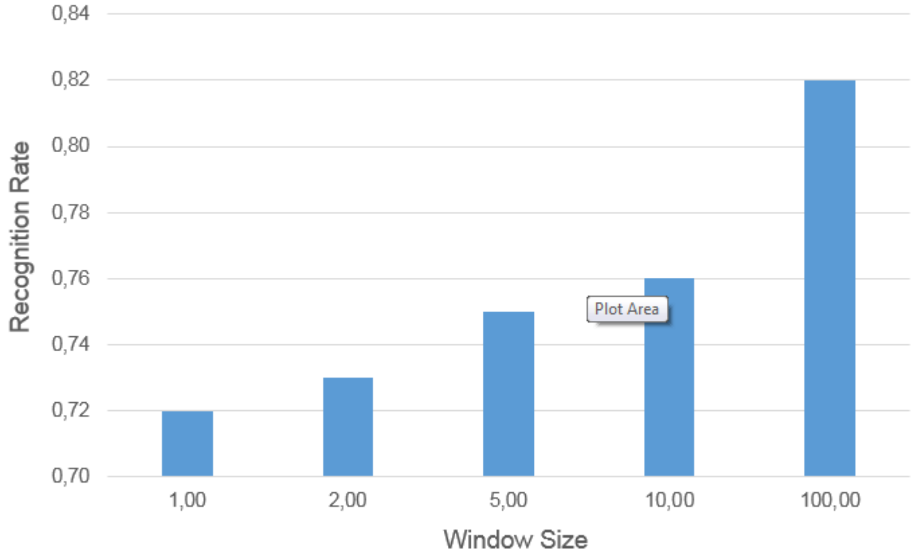


Figure 5.5 – *Effect of the temporal size of the sliding window on low-level feature-based classification on 3D Online Action dataset. The classification rates increase when increasing the length of the temporal window.*

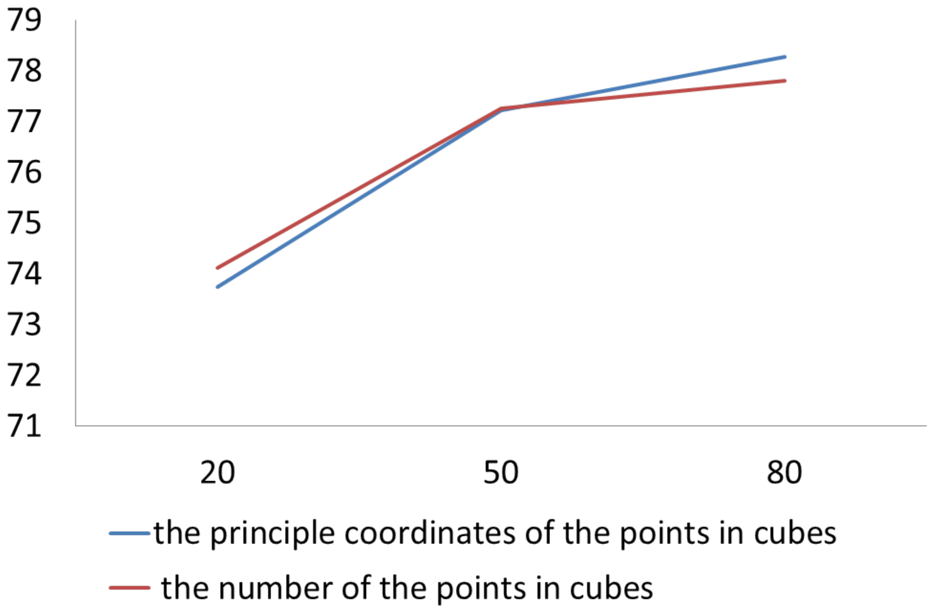


Figure 5.6 – *Effect of the temporal size of the sliding window on object feature-based online classification on 3D Online Action dataset*

Relevant features

We reveal the relevant features for human object interaction recognition. The distances between the object and the joints are selected ones in general. In order to better understand the behavior of the proposed approach, we perform binary classification of each interaction. For action 1 (drinking), we label the data from action 1 as the first class, the second class includes all the remaining actions. The best features to classify action 1 (drinking) are revealed. We repeat this experiments for all the remaining actions separately. Fig. 5.7 shows the results of this experiment. The pairwise distances between the the yellow and red joints are the best features to recognize each human object interaction.

For example, the best features for drinking (action 1) are the pairwise distances between the object joint and the skeleton joints which are on the right hands, on both sides of the crotch, on the left hand and on the left feet. Another example, for eating (action 2), the best features are the pairwise distances between the object joint and the skeleton joints which are on the left hand and on the right hand. There is another situation, for using laptop (action 3), the best features are the pairwise distances between the object joint and the skeleton joints which are on the crotch and on the spinal part. Based on the attributed distances, we know which joint on the skeleton data for each action is more meaningful for recognizing different human object interactions.

MSRDaily Activity 3D Dataset

We will evaluate the MSRDaily Activity 3D dataset[117] for both of online and rate-invariant classifications. It is a daily activity dataset captured by Kinect [69] device, to cover human daily activities in the living room. There are 16 action classes: *drink, eat, read book, call cellphone, write on a paper, use lap- top, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down* each of which was performed twice by 10 subjects. For each video, it provides 3 kinds of data: RGB, depth image and joint and 320 samples in total. Additionally,

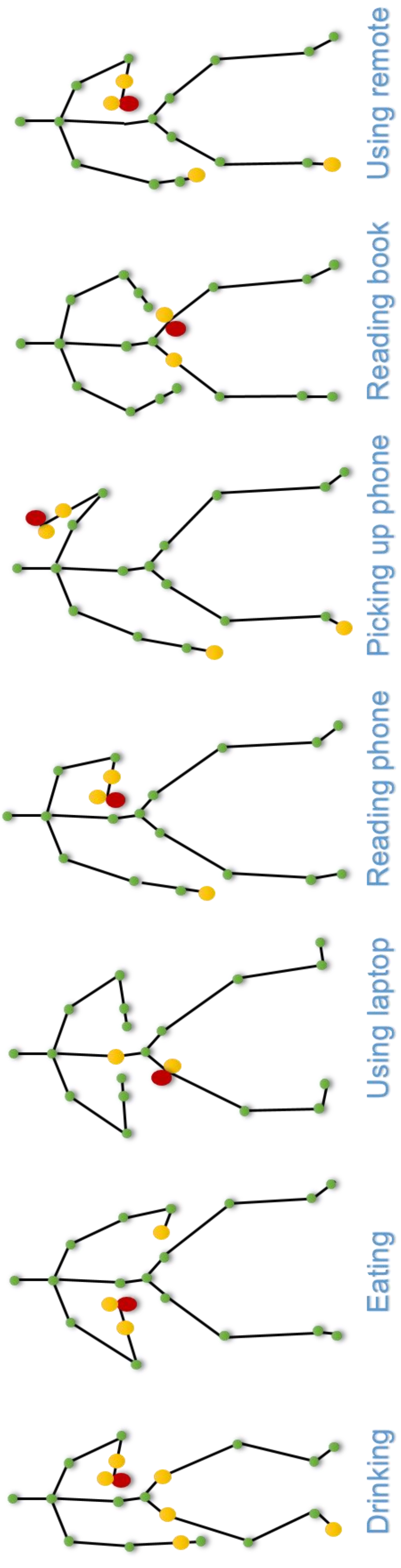


Figure 5.7 – *Selected features for each interaction, the best features are the distances between the yellow and red joints.*

the activities includes human-object interactions and human motion that is the most important reason we choose this dataset.

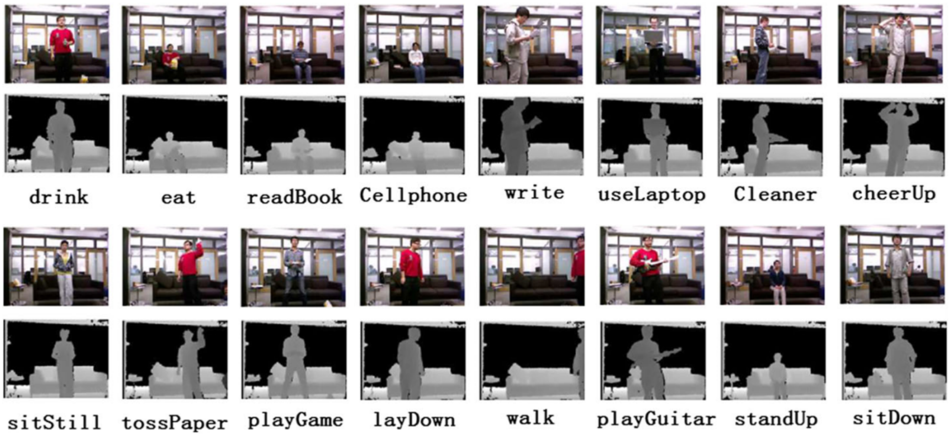


Figure 5.8 – *Selected RGB (top) and raw depth images (bottom) from MSRDaily Activity 3D dataset [117]*

Comparative Evaluation on MSRDaily Activity 3D dataset

In the following we provide a comparative performance analysis of our on-line classification approach with other state-of-the-art solutions using MSRDaily Activity 3D Dataset. We used the same protocol as [139], where half of the subjects are used as training data and the rest of the subjects are used as test data.

Table 5.3 – *Comparison of low-level feature-based online classification on MSRDaily Activity 3D dataset with state of the art results*

Method	Accuracy
[139] (with memory of previous frames)	71.4%
STM of low-level feature	76.35%

As shown in Table 5.3, we compare our method based on low-level feature and object related distance feature with discriminative orderlet mining [139]. To our best knowledge, this is only work on this database for real time classification. We achieves the classification accuracy are 71.4%. The memory both of them are 50 frames.

As we know, most current methods worked on MSRDaily Activity 3D dataset only based on the whole sequences, not online classification. We used the same protocol as [139], where we use the videos from half of the subjects for training and the other half for testing. For fair comparison,

Table 5.4 – *Comparison of object feature-based online classification on MSRDaily Activity 3D dataset with state of the art results*

Method	Accuracy
Continuous Recognition in [139]	60.1%
Number of the points in cubes	74.86%
Principle coordinates of the points in cubes	75.53%

we show our results comparing with online methods in Table 5.4. Here we also adopt two different object features with the same memory set to compare with state of the art results. We can see from Table 5.4, we achieve the classification accuracy of 76.35%, which is much better than the frame level accuracy of discriminative orderlet mining [139]. The best one is the feature based on the principle coordinates of the points in cubes which classification accuracy is 75.53%.

Effect of temporal size of the sliding window on object feature-based online classification

As with 3D Online Action dataset, we can see from Fig 5.9 that there two lines show that: the blue one is the recognition rates of the principle coordinates of the points in cubes on MSRDaily Activity 3D Dataset are respectively 71.59%, 75.33% and 75.53%; the red one is the the recognition rates of the number of the points in cubes on ORGBD dataset are respectively 70.68%, 74.66% and 75.66%.

From the figures, they clearly emerge that the action recognition rate increases when increasing the temporal length of the window. This reveals the importance of the temporal dynamics and shows that the spatio-temporal analysis outperforms a spatial analysis of the frames.

5.3 Off-line human object interaction classification

In this section, the scenario is different here and the challenges also are different. Actually, to recognize a full video, the rate invariance becomes the major challenge.

Several applications require human object interaction recognition after the action is done. This scenario seems less constraining on real-time and

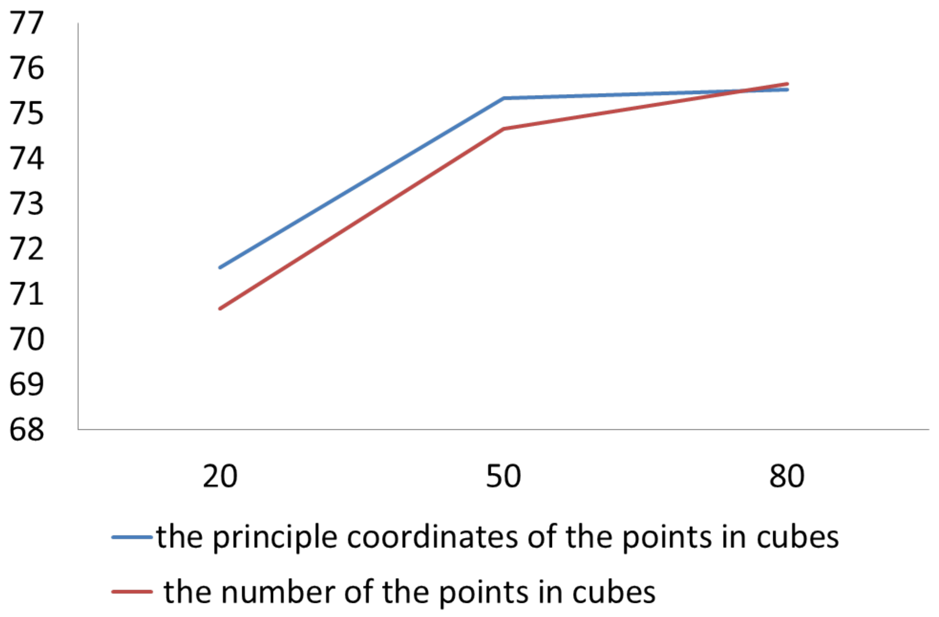


Figure 5.9 – *Effect of temporal size of the sliding window on object feature-based online classification MSRDaily Activity 3D dataset*

rapidity of calculus, however additional challenges arise such as execution time differing for same interaction and significant spatial variation in the way of performing an action.

The object detection and the spatio-temporal modeling are common steps in both scenarios. The input sequences are modeled as trajectories in $\mathbb{R}^{210\times n}$ (where n is the number of frames for each video) via a Spatio-Temporal Modeling (STM). A rate invariant shape analysis of these trajectories is then performed and this make the comparison of the sequences invariant to the rate. The shape analysis framework includes calculation of intrinsic mean of the trajectories issued from the same interaction for training data. The rate invariant distance between a trajectory issued from testing data and all mean trajectories calculated on training data built the final feature vector that represents the input of Random Forest classifier.

In several applications like abnormal activity detection, the earlier the decision can be done the better it is. We propose in this scenario an early detection of human object interaction recognition.

5.3.1 Feature Vector Building

For off-line classification, the final feature vector calculation we use is based on trajectories calculation we introduce in the Chapter 4. So here, we perform our feature vector directly on the MSRDaily Activity 3D dataset including human action and human object interaction for rate invariant classification.

5.3.2 Classification

As the property of rate-invariant classification which is not online method, we just evaluated this method on MSRDaily Activity 3D dataset.

Table 5.5 – *Comparison of Rate-invariant Classification on MSRDaily Activity 3D dataset with state of the art results*

Method	Accuracy
Skeleton in [117]	68.0%
4DHOI model [119]	70.0%
Skeletal shape trajectories [5]	70.0%
Discriminative Orderlet Mining on Batch Recognition [139]	73.8%
Rate-invariant classification	77.05%

As our feature vectors built only based on skeleton joint information, this dataset is very challenging if the depth information is not used. To make it fair for comparison, we mainly compared with the algorithms on skeleton feature [117], [119] and [139]. [5] only used skeleton information that is the same as our work. We used the same experimental setting as [139] and performed on the 2-fold cross-validation which is using the samples of half of the subjects as training data, and the samples of the rest half as testing data. The comparison of the performance is shown in Table 6.1. We can notice in Table 6.1 that we obtained better accuracy than other works. The accuracy of our approach is 77.05%.

Effect of the number of trees in Random Forest algorithm on off-line classification

The performance of Random Forest classifier varies with the number of trees. Thus, we perform the experiments with different numbers of trees; the results of this experimentation is shown in Fig. 5.10. As illustrated in this figure, the recognition rate raises with the increasing number of trees until 60, when the recognition rate reaches 72.5%, and then becomes quite stable. Thus, in the following we consider 50 trees and we report detailed results with this number of trees.

To fully evaluate our method, we perform the experiments with different numbers of trees. So we can see clearly that the performance of Random Forest classifier varies with the number of trees from Fig. 5.10. As illustrated in this figure, the recognition rate raises with the increasing number of trees until 150; the recognition rate reaches the peak 77.05% and then becomes quite stable.

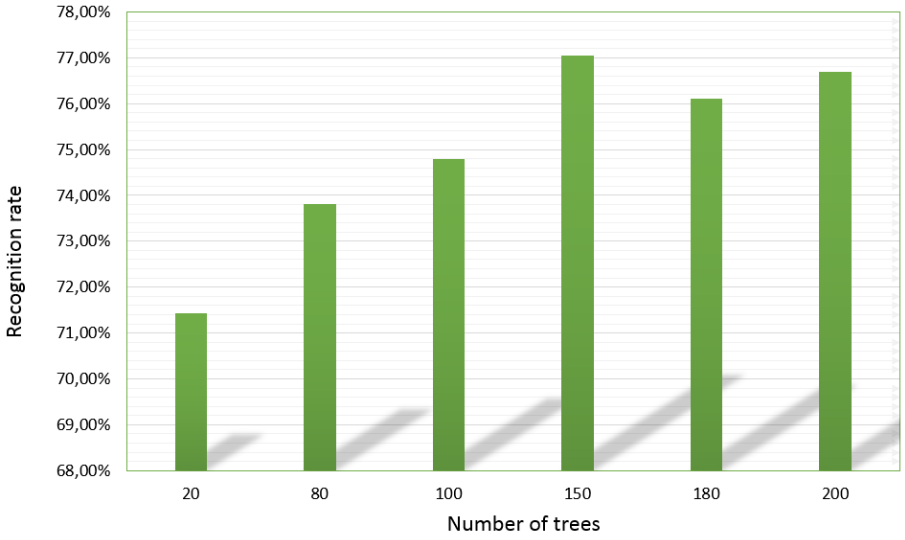


Figure 5.10 – *Human object interaction recognition results using a Random Forest classifier when varying the number of trees.*

5.4 Early recognition

The recognition here is still off-line, however, the earlier the decision can be provided the better it is in several applications like abnormal activities detection. For this end, we provide the recognition rate using the first $k \times 10\%$ of the data, with $k = 1, 2 \ldots 10$. A given test sequence is first modeled as a trajectory in $\mathbb{R}^{210}$ and then the corresponding part of the trajectory is compared to the corresponding part of karcher mean trajectories.

As illustrated in Fig. 5.11, the recognition rate increases slowly using

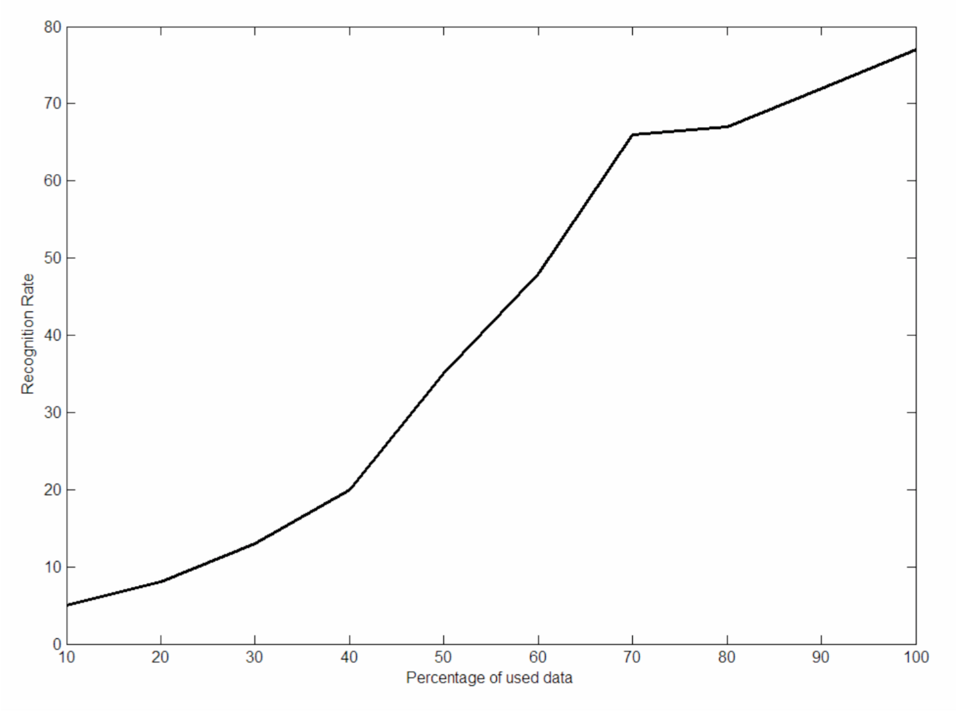


Figure 5.11 – *Early detection: trade off between the recognition rate and the percentage of used data.*

10 to 40% of the data. Then the slope becomes greater until 70% of the data where the performance reaches about 67%. The improvement further is slower using more data.

5.5 Conclusion

In this section, it demonstrated and justified experimentally the effectiveness of the proposed method in the application of human-object interaction recognition within two different scenarios: online recognition and full sequence-based recognition. These two scenarios of human-object interaction recognition reveal different challenges.

For the online recognition, we not only use the set of joints located in the skeleton, the inter-joints distances and the object-joints distances but also object feature which utilized the depth information around two hands were used to classify the human object interaction in real-time.

For the full sequence-based recognition, we use STM to model the pairwise distances of skeleton joints and object joints in each video as a trajectory. Then we compute the mean shape of trajectories corresponding to each action in a rate-invariant way. Human-object interaction classification is solved using Random Forest algorithm applied the feature vector calculated based on the distances to the means of actions.

Experiments performed on MSRDaily Activity 3D dataset and 3D Online dataset testing on human motion and human-object interaction have demonstrated that our proposed approach gives comparative results with respect to state-of-the-art work.

6 Abnormal Activity Recognition

Sommaire

6.1 Introduction

Comparing to verbal or vocal communication data, visual data forms one of the most important cues in developing systems for understanding human behavior. The applications range are from tracking daily activities to classifying emotional states, as well as detecting abnormal and suspicious activities.

As we know, one of the essential measure of human being's performance related to psychological well being is human gait. The character and quality of human gait depends on the strength of the involved muscles and a complex mental coordination process [45]. For this reason, abnormal gait detection is very significant for elderly assistant system, especially the diagnosis of neurological diseases. In addition, the increasing risk of cognitive impairment is affected by physical frailty [74] [113].

So the exploration of deviations from normal gait is necessary for helping current frailty assessment. Besides, a quantitative evaluation of human being's gait attracts more concern for earlier detection of human diseases. In the sense, the present research aims to apply the online framework we used on human object interaction recognition in human abnormal and normal gait analysis.

6.2 Abnormal Activity Recognition

DAI gait dataset

The DAI gait dataset [21] is collected in recording a front view of a corridor which contains seven subjects walk towards the camera normally and abnormal gait. This dataset have two types of anomalies which are knees injured and feet dragged performed by the right and the left leg respectively. So there are four different abnormal gait types. With other four normal instances, each of the seven subjects made of 56 sequences totally. We report some skeleton frames from DGD dataset in Fig. 6.1.

All of the videos are captured by Kinect V2. The experimental setting is the same as in [21] which employed 20 of the available joints, discarding the fingers and a redundant joint of the torso (Spine Shoulder).

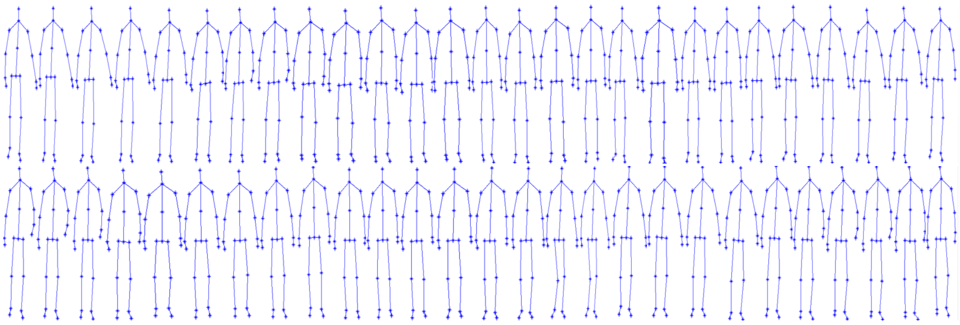


Figure 6.1 – *Some skeleton frames of right knee injured abnormal gait from DAI gait dataset.*

Fig.6.2 shows the joints order of DAI gait dataset.

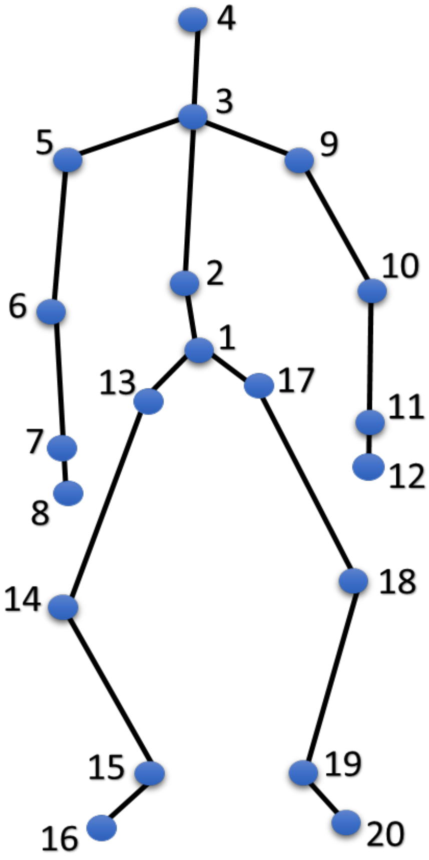


Figure 6.2 – *Joints Order of DAI gait dataset.*

6.2.1 Experimental protocol

The abnormal gait detection is performed on DAI gait dataset following the state-of-the-art protocol. We selected two abnormal and two normal sequences of each subject randomly as training set which contains 28 sequences. The remaining of the dataset is testing set. We compare our approach with the state-of-the-art methods on the cross-subject test setting.

We propose an approach for dynamic abnormal gait detection using frame-based feature and a memory of k previous frames. For $k = 0$, we achieve the recognition without any memory of previous frames. We demonstrate that the use of few distances can be enough for his classification problem.

6.2.2 Experimental results

Table. 6.1 lists the results of our approach and result of the state of the art[21] based on DAI gait dataset. However, our protocol is different from the protocol used in [21]. Actually, in [21], the authors did the training on normal sequences and detect the abnormal ones among the test sequences. Whereas, our training set includes also some abnormal sequences.

Table 6.1 – *Reported results Comparison to state of the art*

Method	Accuracy
Joint Motion History Features (37,42)[21]	98%
Our approach (without memory)	81%
Our approach (with 42 frames)	94.23%
Our approach (with 60 frames)	98.47%

Using a sliding window of size 37 to 42, Alexandros et al. [21] reported 98% F1-measure. We report 81% recognition rate using only one frame, 94.23% using a sliding window of 42 frames and 98.47% using a sliding window of 60 frames.

Effect of the temporal size of the sliding windows

In order to study the effect of size of the sliding window, we report the classification results using several values of the window size as illustrated in Fig. 6.3. One can see that the more frames we use in the window, the better the result is. Using only 37 frames, the recognition rate is 90%. It reaches 94.23% and 97.01% for window size 42 and 50 respectively. Using 60 frames in the sliding window, the recognition rate is 98.47%.

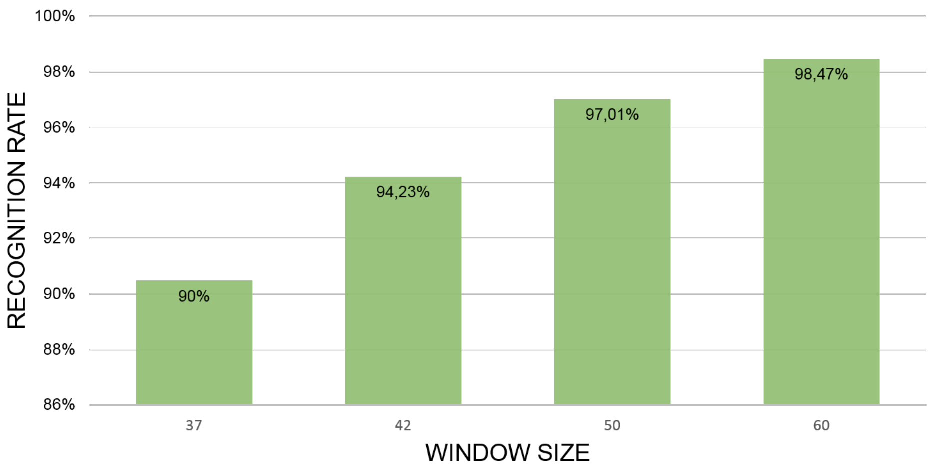


Figure 6.3 – *Effect of the temporal size of the sliding window on the results. The classification rates increase when increasing the length of the temporal window.*

Relevant features

The proposed approach is based on a feature vector of 190 inter joints distances per frame. One important question is are some distances more relevant than others to classify normal and abnormal gaits? Can one do this binary classification (normal versus abnormal) using only one distance?

We investigate these questions and report classification results based on individual distances. We show in Table. 6.2 some classification results based on one distance which has a memory with 40 frames. The distances number 170 is able to classify normal and abnormal gaits with a success of 92.18%. The classification results based on distances 171, 178,188 and 184 are respectively 92%, 89.83%, 93.09% and 97.15%. The distance 179 report a success of 98.46% which is better than the result reported using all the distances together.

Table 6.2 – *Selected features recognition rate*

Number of features	Recognition rate
No.170	92.18%
No.171	92%
No.178	89.83%
No.179	98.46%
No.184	97.15%
No.188	93.09%

We report in Fig. 6.4 and Fig. 6.5 an illustration of the most relevant distances for normal-abnormal gait classification. As illustrated in these

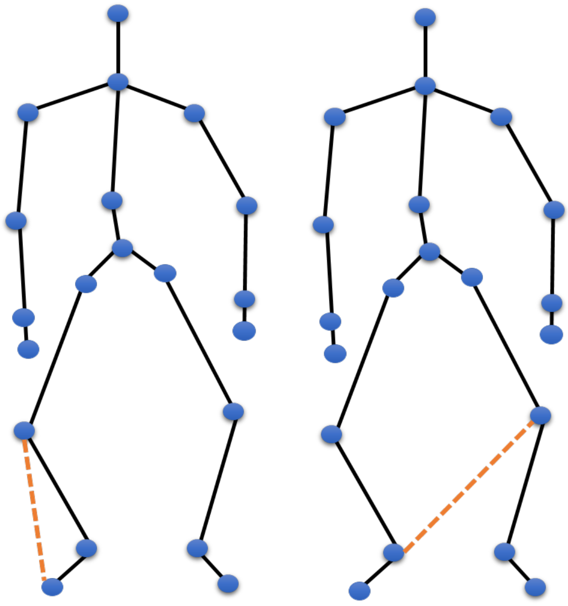


Figure 6.4 – *Illustration of distances No.171 and No.178 from DAI gait Dataset.*

figures, the most relevant distances correspond to the distance between the knee and the foot of the same leg (distance No.171), the distance between knee and the foot of the other leg (distance No.178), the distance between the feet from different legs (distance No.184) and the distance between ankles (distance No.179). This result is in agreement with the data as the 4 abnormal types in the dataset are:

- RKI: Right knee injury (cannot bend the right knee, starting with left foot)
- LKI: Left knee injury (cannot bend the left knee, starting with right foot)
- RFD: Right foot dragging (dragging right leg, starting with left foot)
- LFD: Left foot dragging (dragging left leg, starting with right foot)

In order to better understand the behavior of the relevant distances, we computer the mean normal distance No.179 (distance between right and left ankles) and the mean abnormal one. We show previously that the classification of normal-abnormal gaits using only this distance is better than one based on all distances.

Fig. 6.6 shows the evolution of the mean of this distance in both normal and abnormal cases. One can see that the mean distance is bigger in

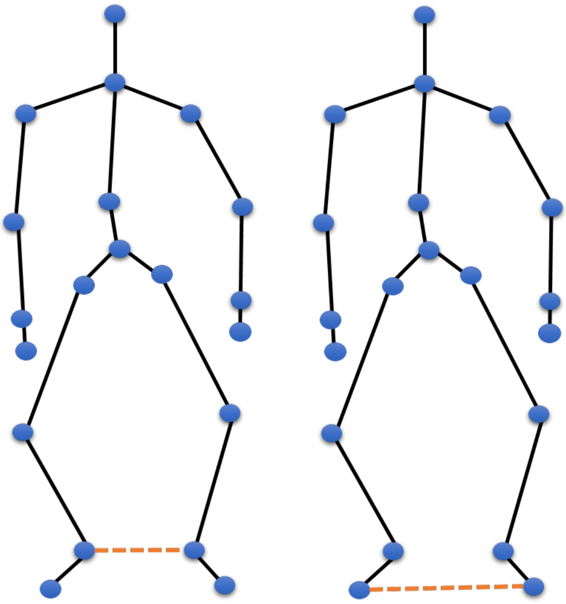


Figure 6.5 – *Illustration of distances No.179 and No.184 from DAI gait Dataset.*

general in the abnormal case and it includes more oscillations than the normal one. It is clear that the variation of this distance is more stable in the normal case.

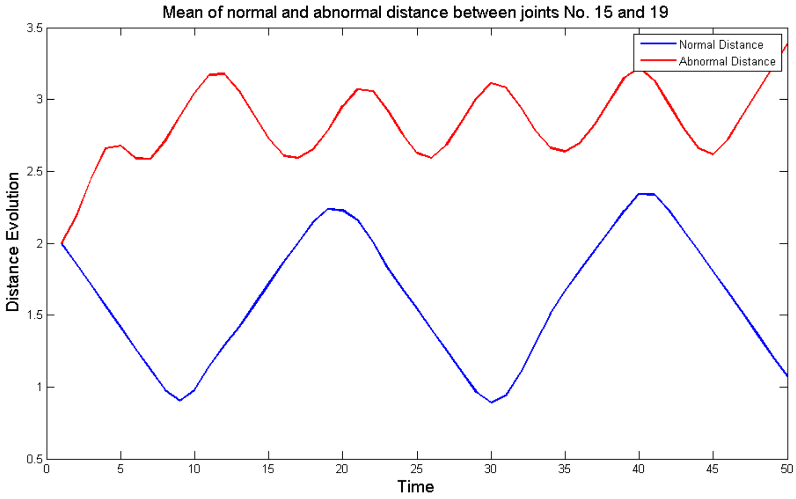


Figure 6.6 – *The mean distance of abnormal and normal gait between joints No.15 and 19.*

Leave-one-actor-out Experiments

We did leave-one-actor-out experiments on DAI gait dataset. For each time, we selected all kinds of actions of one subject as testing set and the rest of the dataset as training set. In total, we have 7 subjects so we did the same experiments seven times. At last, we obtained the mean recognition rate of the seven results. Now we can see the performance of each subject without memory, with 42 frames and with 60 frames as shown in Table. 6.3.

Table 6.3 – *Leave-one-actor-out Experiments*

Subjects	without memory	with 42 frames	with 60 frames
Subject 1	69.9%	88.12%	100%
Subject 2	65.36%	72.99%	88.26%
Subject 3	80.49%	96.58%	100%
Subject 4	66.57%	90.35%	99.12%
Subject 5	77.1%	89.35%	98.43%
Subject 6	77.53%	90.32%	100%
Subject 7	75.41%	93.36%	100%
Mean	73.19%	88.72%	97.97%

From all the results of DAI gait dataset, the proposed approach succeeded in classifying abnormal and normal human actions. The result of experiments reports show that some distances related to the knees, ankles and the feet are more relevant than other distances.

6.3 Multiview 3D human object interaction dataset

To evaluate our method, we built a large-scale 3D event dataset with abnormal and normal human activities involved human object interactions. It is captured by two stationary Kinect sensors from different viewpoints simultaneously. It includes 8 event categories: *press button with injured arm or with injured leg, pick phone with injured arm or with injure leg, use remote and take it back with injured arm or with injure leg, fetch water from dispenser with injured arm, walk around holding cane with injured leg, walk around holding umbrella with injured leg, remove chair with injured leg, walk with plate and put it back on the table with injured arm or with injure leg* and 3 modalities include normal, injured arm and injured leg. All these activities were performed by 10 different subjects each two times in normal and abnormal way. Each event category includes about 30 video sequence instances. For each frame, the Kinect V2 provides 25 skeleton joints which is different from Kinect V1 which provides 20 skeleton joints.

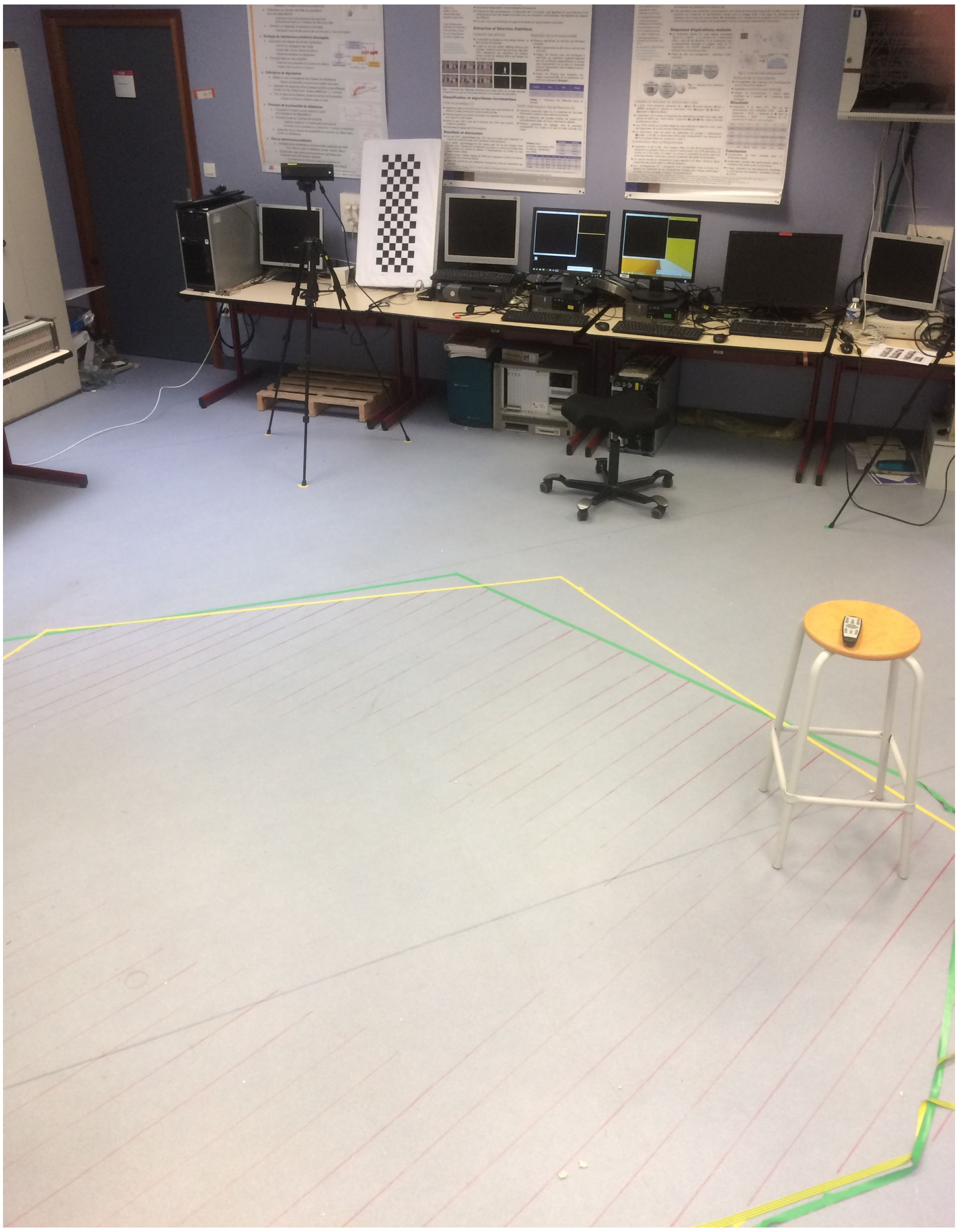


Figure 6.7 – *The setting up of dataset collection*

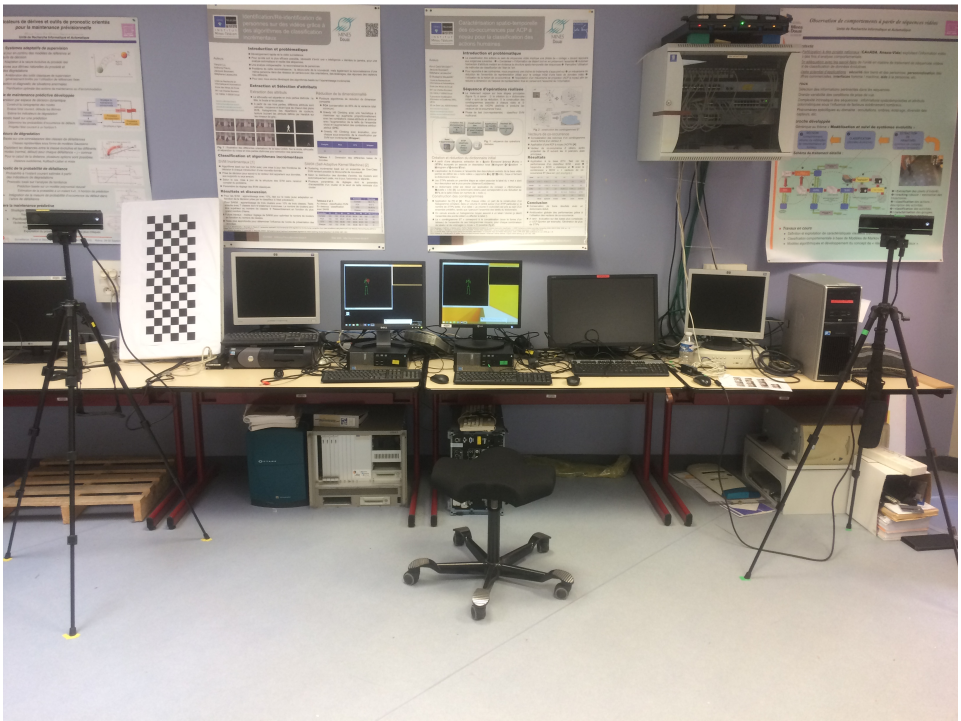


Figure 6.8 – *The setting up of dataset collection*

Here, Fig. 6.7 and Fig. 6.8 are the photographs of the system we set up. We have two Kinects (one on the left and one on the right), mounted on tripods so that we get a big enough common fields (see next forwarded mail to see the trace on the floor (dashed lines) that corresponds to the area where both Kinects provide a good detection of the skeleton. This represent a surface of about 3 x 3 meters starting at (about, again) 1,5 meters from the Kinects' lenses).

There are several characteristics which make the new multi-view dataset challenging. In the first place, we use two Kinect sensors to capture the video. As the various types of subjects' action, the synchronization from different view for rate invariant recognition is a big issue to address. In the second place, we not only capture the normal persons holding different objects but also abnormal persons executing activities with objects. At last, there are two abnormal modalities which means our new dataset has large variety when each subject performing an event.

6.3.1 Experimental protocol

Due to the new multi-view data, we test our on-line framework on the new dataset in two main scenarios: different views and synchronized view. In the scenario of synchronized view, we divide it into two different experiments based on person independent or not. For each scenario, there are three different protocols according to the properties of the new dataset. In the first protocol, there are two classes which are normal and abnormal. As there are two types of abnormal modalities, we make three classes: normal, injured arm and injured leg in the second protocol. In the third protocol, we class them by different types of activities.

So in the experiments of different views, we use all of the videos from the one of Kinect sensors as the training set and all of the videos from the other. In the experiments of synchronized view on person dependent, due to each action performing twice, we use all the first iteration videos as training set and the second time as testing set. In addition, for synchronized view on person independent, we choose all the videos from half actors as training set and the other half of actors' videos as testing data. All these experiments were based on the 2-fold cross-validation.

Trajectories synchronization

For the synchronized view experiments, we propose two steps for synchronizing the same action from different views. First, we use the resampling algorithm and the function 4.2 of shape analysis framework for the alignment of trajectories from each frame. Second, after alignment, we adopt the proposed fusion framework to achieve a fused trajectories by selecting the best distance attributes from each trajectory. For the fusion algorithm, we will detail it later.

As shown in Fig. 6.9, we apply the same framework like we did in the temporal modeling. The β_1 and β_2 refer to the feature matrix obtained from the low-level feature based on the same action from different views. $\beta_1(t_0)$ is the trajectory of first frame of β_1 and $\beta_1(t_1)$ is the trajectory of second frame of β_1. So the $\beta_1(t)$ is the one of trajectories of the frames of

β_1 and β_2 is the same. So after resampling, we align them frame by frame.

$$\beta_2 * \gamma^*$$

is the aligned trajectory after applying the function γ^* (4.2). Then, we fuse the β_1 and $\beta_2 * \gamma^*$ as following for the same action.

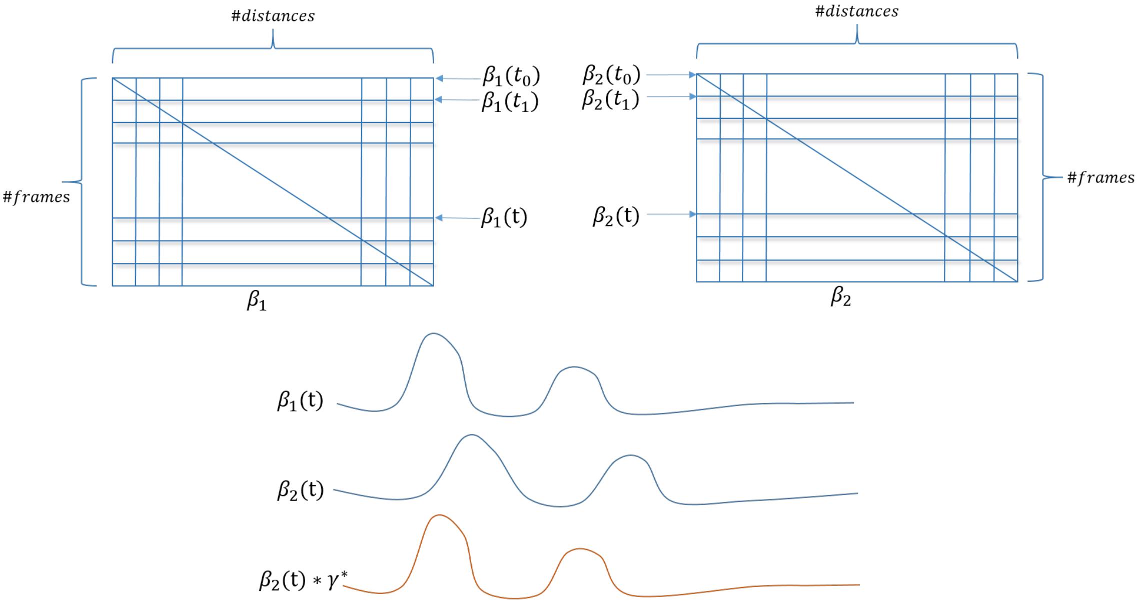


Figure 6.9 – *The illustration of trajectories synchronization*

Trajectories fusion

As explained above, we need fuse the trajectories for each frame of the videos from K1 and K2 which refer to these two Kinect sensors. In Fig. 6.10, we can notice that there are three colors to represent different distance attributes. Here, we calculate the distance between $\beta_1(t)$ and $\beta_2(t)$: if they are not close, we choose the mean of the corresponding distances (red); if they are close, we choose the smoother one by comparing their own curvature 6.1.

$$k = |dt/ds| \tag{6.1}$$

t is the velocity vector which is also the difference between the distance attributes on $\beta_1(t)$ and $\beta_2(t)$.

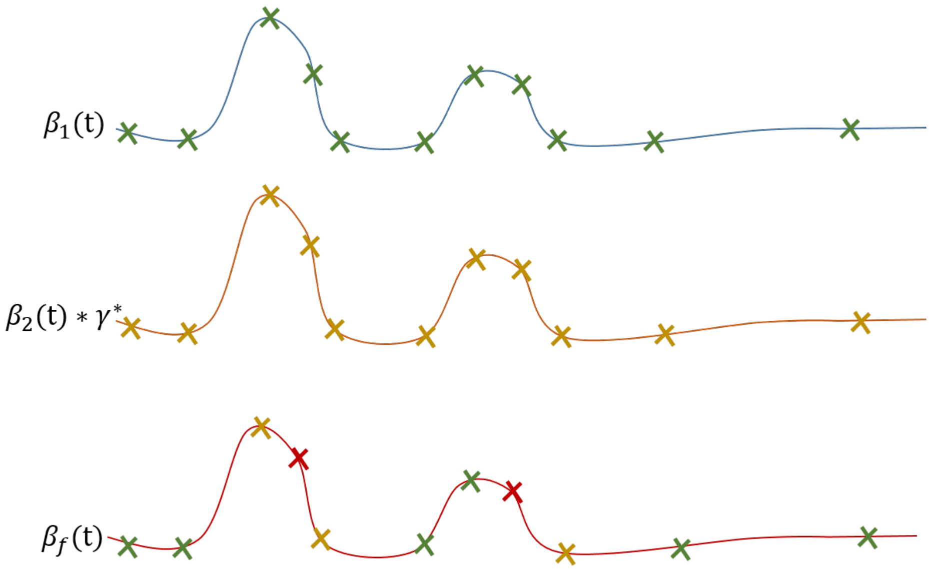


Figure 6.10 – *The illustration of trajectories fusion. The green, yellow and red points refer to the distance on $\beta(t)$ of the videos from K1, K2 after synchronization and mean respectively.*

6.3.2 Experimental results

Table 6.4 – *The results of different scenarios for the task of multi-view human object interaction recognition*

Protocols	Accuracy (%)
Different views	
Protocol1	74.32
Protocol2	60.61
Protocol3	77.05
Synchronized view on person dependent	
Protocol1	67.12
Protocol2	55.24
Protocol3	70.17
Synchronized view on person independent	
Protocol1	62.21
Protocol3	51.41
Protocol3	65.27

During these experiments, we built our feature vectors based on the on-line framework. Due to the different characters of two scenarios, different views experiments use the low-level feature for the task of abnormal and normal human object interaction recognition. In the synchronized view experiments, we build the feature vectors based on fusing the low-level feature of the same action from different view by shape analysis framework.

By analyzing Table. 6.4, it can be noticed that the results of the two scenarios based on the three protocols show the success of the proposed method.

6.4 Conclusion

In this section, we propose a spatio-temporal modeling of the skeleton data based on inter-joint distances for normal and abnormal gaits classification. The proposed features are discriminative enough to classify abnormal and normal human actions. We report 98.47% recognition rate and show that some distances related to the knees, ankles and the feet are more relevant than other distances. Future work will be focused on more features to be able to distinguish the different types of abnormal gaits.

In addition, we collected a RGB-D-based multi-view 3D human object interaction dataset including abnormal and normal human behaviors by two Kinect sensors. We test our model on the new dataset by two different scenarios: different views and synchronized view. The evaluation on the scenario of different views obtained best recognition rate which is 77.05% and also proved the effectiveness of the proposed framework.

7 Conclusion

Sommaire

7.1 Summary

In this thesis, we presented contributions to the human activity recognition using low-cost 3D sensors with a focus on human-object interactions. We demonstrated and justified experimentally the effectiveness of our method in two main applications; the abnormal gait detection and human-object interaction recognition within two different scenarios: online recognition and full sequence-based off-line recognition. The last two scenarios of human-object interaction recognition reveal different challenges.

We model activities in spatial and temporal way respectively for different applications. Firstly, we propose low-level feature is composed of inter-joints distance and object related distance which adopted on online human object interaction recognition and abnormal gait detection.

Secondly, we propose object feature, a rough description of the object shape and size as new features to model the human-object interactions. This object feature is fused with the low-level feature for online human activity recognition. These features have the advantage to be pose, position invariant and discriminative to model the human articulations movement in a given frame. A Random Forest based classification algorithm to this end.

The experiments about online human object interaction recognition conducted on 3D Online dataset and MSRDaily Activity dataset respectively, following state-of-the-art setting demonstrate the effectiveness of the proposed framework. The abnormal gait detection is performed on DAI gait dataset following the state-of-the-art protocol and show that the proposed approach success to classify abnormal and normal human actions. We report 98.47% recognition rate and show that some distances related to the knees, ankles and the feet are more relevant than other distances.

Furthermore, the full sequence-based human object interaction recognition seems easier scenario compared to online recognition. However, this scenario reveals a new challenge which consists on rate invariance. A more elaborated spatio-temporal modeling is proposed here. The evolution of the inter-joints and object-joints distances in time is modeled as trajectories in a high dimension space and a shape analysis framework is used to analyze and compare the corresponding trajectories in a Riemannian manifold. This framework has the advantage to make the reparameterization group acting by isometry on the space of these trajectories. The distance between the orbits corresponding to two trajectories is invariant to the rate of execution in the sequence. Another advantage of the used shape analysis framework is the calculation of intrinsic means which are rate invariant. This helps to summarize the shape of trajectories belonging to the same class and accelerates the classification based on Random Forest.

Experiments performed on MSRDaily Activity dataset have demonstrated that our proposed approach gives comparative results with respect to state-of-the-art work.

To evaluate our algorithm, we built a multi-view 3D human object interaction dataset including abnormal and normal human activities using two simultaneous Kinect sensors from different viewpoints around the subjects. The experiment results on this dataset show the effectiveness of our method.

7.2 Limitations and future work

In this section, we briefly describe some directions that could extend our work. As mentioned before, it is necessary to deal with the issues like segmentation, modeling, and occlusion handling. Beyond that, the humanoid image model representations are high sensitive to the inter-class activities. So an invariant system exploration is one of major challenges for same human object interaction class with wide variability in the features.

Except action execution rate we deal with in this thesis, camera view point is significant for applications such as surveillance. There are some successful approaches on single view but for multi-view miserably. So there is the need to explore a framework for a given action captured from the different viewpoints for different people.

Additionally, intention reasoning plays a very important role in security applications. For most of the cases, the system is trained for specific actions, and may fail for different unpredicted actions. For the detection of fighting, if the system is trained for kicking and punching, it may fail to detect hitting with an object in fighting even though it is a part of fighting. Today's systems may fail to identify the difference between Karate practice and actual fighting due to ignorance of the intention reasoning parameter. So, there is a strong need of designing robust generalized systems with intention reasoning.

Finally, providing product for real world problems is the most important issue we need address in future work. Various available algorithms for human motion representation, and recognition are mainly, driven by specific applications or datasets. Many researchers and organizations are actively involved in the domain, and have provided a variety of datasets. In order to design and deploy vision based systems for various surveillance, and control and analysis applications in real time environment, there is the need of a rigorously prepared, standardized common dataset to assess and compare the algorithmic performances. Advances in other domain can be applied for more accurate results. The future work should lead the Computer Vision community to provide a robust solution for real world problems in various applications.

Bibliography

[1] Dib Abdallah and François Charpillet. Pose estimation for a partially observable human body from rgb-d cameras. In *IEEE/RJS International Conference on Intelligent Robots and Systems (IROS)*, page 8, 2015. (Cited page 21.)

[2] M. F. Abdelkader, W. Abd-Almageed, A. Srivastava, and R. Chellappa. Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. *Computer Vision and Image Understanding*, 115(3):439–455, 2011. (Cited page 25.)

[3] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011. (Cited pages 4, 11, and 25.)

[4] E. E. Aksoy, A. Abramov, J. Dorr, K. Ning, B. Dellen, and F. Worgotter. Learning the semantics of objectaction relations by observation. *The International Journal of Robotics Research*, 0278364911410459, 2011. (Cited page 26.)

[5] B. B. Amor, J. Su, and A. Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38(1):1–13, 2016. (Cited page 80.)

[6] Boulbaba Ben Amor, Jingyong Su, and Anuj Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):1–13, 2016. (Cited page 35.)

[7] R. Anirudh, P. Turaga, J. Su, and A. Srivastava. Elastic functional coding of human actions: from vector-fields to latent variables. *IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 3147–3155, 2015. (Cited pages 4 and 25.)

[8] ASUS Xtion PRO LIVE. `http://www.asus.com/Multimedia/Xtion_PRO/`, 2013. (Cited page 16.)

[9] Lin Bai, Kan Li, Jianmeng Pei, and Shuai Jiang. Main objects interaction activity recognition in real images. *Neural Computing and Applications*, 27(2):335–348, 2016. (Cited page 28.)

[10] G. E. Batista and M. C. Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17:519–533, 2003. (Cited page 47.)

[11] Marco Bertini, Alberto Del Bimbo, and Walter Nunziati. Model checking for detection of sport highlights. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 215–222. ACM, 2003. (Cited page 15.)

[12] Victoria Bloom, Vasileios Argyriou, and Dimitrios Makris. G3di: A gaming interaction dataset with a real time detection and evaluation framework. In *Workshop at the European Conference on Computer Vision*, pages 698–712. Springer, 2014. (Cited pages ix, 21, and 24.)

[13] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Unsupervised feature learning for rgb-d based object recognition. In *Experimental Robotics*, pages 387–402. Springer, 2013. (Cited page 18.)

[14] Ricard Borràs, Àgata Lapedriza, and Laura Igual. Depth information in human gait analysis: an experimental study on gender recognition. In *International Conference Image Analysis and Recognition*, pages 98–105. Springer, 2012. (Cited page 21.)

[15] Oliver Brdiczka, Matthieu Langet, Jérôme Maisonnasse, and James L Crowley. Detecting human behavior models from multimodal observation in a smart home. *IEEE Transactions on Automation Science and Engineering*, 6(4):588–597, 2009. (Cited page 13.)

[16] L. Breiman. Machine learning. 45:5–32, 2001. (Cited page 66.)

[17] Elizabeth Broadbent, Rebecca Stafford, and Bruce MacDonald. Acceptance of healthcare robots for the older population: review and future directions. *International Journal of Social Robotics*, 1(4):319–330, 2009. (Cited page 13.)

[18] Xingyang Cai, Wengang Zhou, Lei Wu, Jiebo Luo, and Houqiang Li. Effective active skeleton representation for low latency human action recognition. *IEEE Transactions on Multimedia*, 18(2):141–154, 2016. (Cited page 35.)

[19] Ziyun Cai, Jungong Han, Li Liu, and Ling Shao. Rgb-d datasets using microsoft kinect or similar sensors: a survey. *Multimedia Tools and Applications*, pages 1–43. (Cited page 20.)

[20] Jiaya Jia Cewu Lu and Chi-Keung Tang. Range-sample depth feature for action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Columbus, USA, june 2014. (Cited page 31.)

[21] Alexandros Andre Chaaraoui, José Ramón Padilla-López, and Francisco Flórez-Revuelta. Abnormal gait detection with rgb-d devices using joint motion history features. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 7, pages 1–6. IEEE, 2015. (Cited pages x, 20, 39, 40, 86, and 88.)

[22] J. M. Chaquet, E. J. Carmona, and A. FernÃ¡ndez-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013. (Cited page 25.)

[23] Chen Chen, Roozbeh Jafari, and Nasser 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*, pages 168–172. IEEE, 2015. (Cited page 21.)

[24] G. Chen, D. Clarke, M. Giuliani, A. Gaschler, and A. Knoll. Combining unsupervised learning and discrimination for 3d action recognition. *Signal Processing*, 110:67–81, 2015. (Cited page 25.)

[25] Lulu Chen, Hong Wei, and James Ferryman. A survey of human motion analysis using depth imagery. *Pattern Recognition Letters*, 34(15):1995–2006, 2013. (Cited page 18.)

[26] X Chen, Mark Billinghurst, SA Green, and GJ Chase. Human-robot collaboration: A literature review and augmented reality approach in design. 2008. (Cited page 13.)

[27] TT Dao, H Tannous, P Pouletaut, D Gamet, D Istrate, and MC Ho Ba Tho. Interactive and connected rehabilitation systems for e-health. *IRBM*, 2016. (Cited page 15.)

[28] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros. Scene semantics from long-term observation of people. *European Conference on Computer Vision (ECCV)*, pages 248–298, 2012. (Cited page 26.)

[29] Emel Demircan, Dana Kulic, Denny Oetomo, and Mitsuhiro Hayashibe. Human movement understanding [tc spotlight]. *IEEE Robotics & Automation Magazine*, 22(3):22–24, 2015. (Cited page 13.)

[30] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. *European Conference on Computer Vision (ECCV)*, 2012. (Cited page 26.)

[31] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 95(1):1–12, 2011. (Cited page 26.)

[32] M. Devanne, H. Wannous, P. Berretti, S.and Pala, M. Daoudi, and A Del Bimbo. 3d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Transactions on Cybernetics*, 45(7):1340–1352, 2015. (Cited page 35.)

[33] M. Devanne, H. Wannous, P. Berretti, S.and Pala, M. Daoudi, and A Del Bimbo. Combined shape analysis of human poses and motion units for action segmentation and recognition. *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FGW)*, 7:1–6, 2015. (Cited page 35.)

[34] Wenwen Ding, Kai Liu, Fei Cheng, and Jin Zhang. Learning hierarchical spatio-temporal pattern for human activity prediction. *Journal of Visual Communication and Image Representation*, 35:103–111, 2016. (Cited pages x, 35, and 36.)

[35] Chris Ellis, Syed Zain Masood, Marshall F Tappen, Joseph J Laviola Jr, and Rahul Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision*, 101(3):420–436, 2013. (Cited page 21.)

[36] Masahiro Fujita. Digital creatures for future entertainment robotics. In *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on*, volume 1, pages 801–806. IEEE, 2000. (Cited page 13.)

[37] Juergen Gall, Andrea Fossati, and Luc Van Gool. Functional categorization of objects using real-time markerless motion capture. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1969–1976. IEEE, 2011. (Cited pages x and 30.)

[38] Samuele Gasparrini, Enea Cippitelli, Susanna Spinsante, and Ennio Gambi. A depth-based fall detection system using a kinect® sensor. *Sensors*, 14(2):2756–2775, 2014. (Cited page 21.)

[39] Weina Ge, Robert T Collins, and R Barry Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):1003–1016, 2012. (Cited page 13.)

[40] Dian Gong and Gerard Medioni. Dynamic manifold warping for view invariant action recognition. In *2011 International Conference*

on Computer Vision, pages 571–578. IEEE, 2011. (Cited pages x, 32, and 33.)

[41] Hans W Guesgen and Stephen Marsland. Using contextual information for recognising human behaviour. *International Journal of Ambient Computing and Intelligence (IJACI)*, 7(1):27–44, 2016. (Cited page 14.)

[42] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human- object interactions: using spatial and functional compatibility for recognition. *The International Journal of Robotics Research*, 31(10):1775–1789, 2009. (Cited page 26.)

[43] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1961–1968. IEEE, 2011. (Cited pages x, 32, and 33.)

[44] Fei Han, Brian Reily, William Hoff, and Hao Zhang. space-time representation of people based on 3d skeletal data: a review. *arXiv preprint arXiv:1601.01006*, 2016. (Cited pages 36 and 66.)

[45] Jeffrey M Hausdorff, Galit Yogev, Shmuel Springer, Ely S Simon, and Nir Giladi. Walking is more like catching than tapping: gait in the elderly as a complex cognitive task. *Experimental Brain Research*, 164(4):541–548, 2005. (Cited page 86.)

[46] Du Q Huynh Hossein Rahmani, Arif Mahmood and Ajmal Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European Conference on Computer Vision (ECCV)*, Zurich, Swiss, September 2014. (Cited page 31.)

[47] Gibson Hu, Shoudong Huang, Liang Zhao, Alen Alempijevic, and Gamini Dissanayake. A robust rgb-d slam algorithm. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1714–1719. IEEE, 2012. (Cited page 18.)

[48] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recogni-

tion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5344–5352, 2015. (Cited page 21.)

[49] Dong Huang, Shitong Yao, Yi Wang, and Fernando De La Torre. Sequential max-margin event detectors. In *European conference on computer vision*, pages 410–424. Springer, 2014. (Cited page 21.)

[50] M. Jiang, J. Kong, G. Bebis, and H. Huo. Informative joints based human action recognition using skeleton contexts. *Signal Processing: Image Communication*, 33:29–40, 2015. (Cited page 25.)

[51] Bongjin Jun, Inho Choi, and Daijin Kim. Local transform features and hybridization for accurate face and human detection. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1423–1436, 2013. (Cited page 13.)

[52] Geetanjali Kale and Varsha Patil. Bharatnatyam adavu recognition from depth data. In *2015 Third International Conference on Image Information Processing (ICIIP)*, pages 246–251. IEEE, 2015. (Cited page 14.)

[53] Geetanjali Vinayak Kale and Varsha Hemant Patil. A study of vision based human motion recognition and analysis. *International Journal of Ambient Computing and Intelligence (IJACI)*, 7(2):75–92, 2016. (Cited page 13.)

[54] Kyong Il Kang, Sanford Freedman, Maja J Mataric, Mark J Cunningham, and Becky Lopez. A hands-off physical therapy assistance robot for cardiac patients. In *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.*, pages 337–340. IEEE, 2005. (Cited page 13.)

[55] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977. (Cited page 36.)

[56] S. R. Ke, H. L. U. Thuc, Y. J. Lee, J. N. Hwang, J. H. Yoo, and K. H. Choi. A review on video-based human activity recognition. *Computers*, 2(2):88–131, 2013. (Cited page 25.)

[57] H. Kjellstrom, J. Romero, and D. Kragic. Visual object-action recognition: inferring object affordances from human demonstration. *International Journal of Computer Vision*, 115(1):81–90, 2011. (Cited page 26.)

[58] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33, 2013. (Cited pages 10 and 19.)

[59] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. (Cited page 36.)

[60] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. (Cited pages ix, 21, and 22.)

[61] Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer methods and programs in biomedicine*, 117(3):489–501, 2014. (Cited page 21.)

[62] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 9–14. IEEE, 2010. (Cited pages 29 and 30.)

[63] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 812–819, 2014. (Cited page 21.)

[64] Jonathan Feng-Shun Lin and Dana Kulić. Online segmentation of human motion for automated rehabilitation exercise analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):168–180, 2014. (Cited page 15.)

[65] Li Liu and Ling Shao. Learning discriminative representations from rgb-d video data. In *IJCAI*, volume 1, page 3, 2013. (Cited page 18.)

[66] Dimitrios Makris and Tim Ellis. Learning semantic scene models from observing activity in visual surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(3):397–408, 2005. (Cited page 14.)

[67] Arun Mallya and Svetlana Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. *arXiv preprint arXiv:1604.04808*, 2016. (Cited pages ix and 28.)

[68] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2929–2936, 2009. (Cited pages 25 and 26.)

[69] Microsoft. Microsoft kinect. *http://www.microsoft.com/en-us/ kinectfor-windows/*, 2013. (Cited pages 16 and 75.)

[70] Microsoft Corporation Kinect v2 for Xbox 360. `http://www.xbox.com/en-GB/xbox-one/accessories/`, 2016. (Cited page 16.)

[71] Microsoft Kinect Software Development Toolkit. `https://dev.windows.com/en-us/kinect/develop/`, 2016. (Cited page 16.)

[72] Francesco Mondada, Michael Bonani, Xavier Raemy, James Pugh, Christopher Cianci, Adam Klaptocz, Stephane Magnenat, Jean-Christophe Zufferey, Dario Floreano, and Alcherio Martinoli. The e-puck, a robot designed for education in engineering. In *Proceedings of the 9th conference on autonomous robot systems and competitions*, volume 1, pages 59–65. IPCB: Instituto Politécnico de Castelo Branco, 2009. (Cited page 13.)

[73] D. J. Moore, I. A. Essa, and M. H. Hayes III. Exploiting human actions and object context for recognition tasks. *IEEE Conference on Computer Vision (ICCV)*, 1:80–81, 1999. (Cited page 26.)

[74] Natalia E Morone, Kaleab Z Abebe, Lisa A Morrow, and Debra K Weiner. Pain and decreased cognitive function negatively impact

physical functioning in older adults with knee osteoarthritis. *Pain Medicine*, 15(9):1481–1487, 2014. (Cited page 86.)

[75] Brendan Tran Morris and Mohan Manubhai Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE transactions on circuits and systems for video technology*, 18(8):1114–1127, 2008. (Cited page 14.)

[76] Bijan Najafi, Kamiar Aminian, Anisoara Paraschiv-Ionescu, François Loew, Christophe J Bula, and Philippe Robert. Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. *IEEE Transactions on biomedical Engineering*, 50(6):711–723, 2003. (Cited page 15.)

[77] Farhood Negin, Fırat Özdemir, Ceyhun Burak Akgül, Kamer Ali Yüksel, and Aytül Erçil. A decision forest based feature selection framework for action recognition from rgb-depth cameras. In *International Conference Image Analysis and Recognition*, pages 648–657. Springer, 2013. (Cited page 21.)

[78] Bingbing Ni, Gang Wang, and Pierre Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013. (Cited pages ix, 21, 23, and 24.)

[79] Bingbing Ni, Xiaokang Yang, and Shenghua Gao. Progressively parsing interactional objects for fine grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1020–1028, 2016. (Cited pages ix and 29.)

[80] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BmVC*, volume 1, page 3, 2011. (Cited page 18.)

[81] Kei Okada, Takashi Ogura, Atsushi Haneda, Junya Fujimoto, Fabien Gravot, and Masayuki Inaba. Humanoid motion generation system on hrp2-jsk for daily life environment. In *IEEE International Confer-*

ence Mechatronics and Automation, 2005, volume 4, pages 1772–1777. IEEE, 2005. (Cited page 13.)

[82] OpenNI 2 Software Development Toolkit. `http://structure.io/openni/`, 2016. (Cited page 16.)

[83] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 716–723, 2013. (Cited pages 21, 30, and 32.)

[84] Benjamin Packer, Kate Saenko, and Daphne Koller. A combined pose, object, and feature model for action understanding. In *CVPR*, pages 1378–1385. Citeseer, 2012. (Cited page 30.)

[85] Adeline Paiement, Lili Tao, Sion Hannuna, Massimo Camplani, Dima Damen, and Majid Mirmehdi. Online quality assessment of human movement from skeleton data. In *Proceedings of British Machine Vision Conference*, 2014. (Cited pages x and 34.)

[86] Gemma S Parra-Dominguez, Babak Taati, and Alex Mihailidis. 3d human motion analysis to detect abnormal events on stairs. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 97–103. IEEE, 2012. (Cited page 39.)

[87] H. Pazhoumand-Dar, C. P. Lam, and M Masek. Joint movement similarities for robust 3d action recognition using skeletal data. *Journal of Visual Communication and Image Representation*, 30:10–21, 2015. (Cited page 25.)

[88] A. Prest, V. Ferrari, and C. Schmid. Explicit modeling of human-object interactions in realistic videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(4):835–848, 2013. (Cited page 26.)

[89] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence (PAMI), 34(3):601–614, 2012. (Cited page 27.)

[90] Jin Qi and Zhiyong Yang. Learning dictionaries of sparse codes of 3d movements of body joints for real-time human activity understanding. *PloS one*, 9(12):e114147, 2014. (Cited page 35.)

[91] Bellman R. and Dreyfus S. Applied dynamic programming. *RAND Corporation*, 1962. (Cited page 127.)

[92] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European Conference on Computer Vision*, pages 742–757. Springer, 2014. (Cited page 21.)

[93] Marco Ronchetti and Mattia Avancini. Using kinect to emulate an interactive whiteboard. *MS in Computer Science, University of Trento*, 2011. (Cited page 15.)

[94] C. Schmid S. Lazebnik and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, New York, USA, June 2006. (Cited page 31.)

[95] Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE transactions on pattern analysis and machine intelligence*, 27(2):162–177, 2005. (Cited page 14.)

[96] Afshin Sepehri, Yaser Yacoob, and Larry S Davis. Employing the hand as an interface device. *Journal of Multimedia*, 1(7):18–29, 2006. (Cited page 15.)

[97] Amir Shahroudy, Tian-Tsong Ng, Qingxiong Yang, and Gang Wang. Multimodal multipart learning for action recognition in depth videos. *IEEE transactions on pattern analysis and machine intelligence*, 2015. (Cited pages x, 37, and 38.)

[98] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, pages 116–124, 2013. (Cited pages 18, 29, and 68.)

[99] R. Slama, H. Wannous, M. Daoudi, and A. =Srivastava. Accurate 3d action recognition using learning on the grassmann manifold. *Pattern Recognition*, 48(2):556–567, 2015. (Cited pages 4 and 25.)

[100] Jasper Snoek, Jesse Hoey, Liam Stewart, Richard S Zemel, and Alex Mihailidis. Automated detection of unusual events on stairs. *Image and Vision Computing*, 27(1):153–166, 2009. (Cited pages x, 38, and 39.)

[101] Yale Song, David Demirdjian, and Randall Davis. Continuous body and hand gesture recognition for natural human-computer interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(1):5, 2012. (Cited page 13.)

[102] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(7):1415–1428, 2011. (Cited pages 25, 56, 122, 123, and 126.)

[103] Anuj Srivastava, Eric Klassen, Shantanu H. Joshi, and Ian H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *PAMI*, 33(7):1415–1428, 2011. (Cited pages 58 and 59.)

[104] J. Su, A. Srivastava, F. Souza, and S. Sarkar. Rate-invariant analysis of trajectories on riemannian manifolds with application in visual speech recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 620–627, 2014. (Cited page 35.)

[105] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Human activity detection from rgbd images. *plan, activity, and intent recognition*, 64, 2011. (Cited page 21.)

[106] Ilias Theodorakopoulos, Dimitris Kastaniotis, George Economou, and Spiros Fotopoulos. Pose-based human action recognition via

sparse representation in dissimilarity space. *Journal of Visual Communication and Image Representation*, 25(1):12–23, 2014. (Cited page 21.)

[107] Emanuele Trucco and Alessandro Verri. *Introductory techniques for 3-D computer vision*, volume 201. Prentice Hall Englewood Cliffs, 1998. (Cited page 17.)

[108] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(11):2273–2286, 2011. (Cited pages 4 and 25.)

[109] Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008. (Cited pages 13 and 44.)

[110] Md Zia Uddin. Human activity recognition using segmented body part and body joint features with hidden markov models. *Multimedia Tools and Applications*, pages 1–30, 2016. (Cited pages 37 and 38.)

[111] A. Veeraraghavan, A. Srivastava, A. K. Roy-Chowdhury, and R. Chellappa. Rate-invariant recognition of humans and their activities. *IEEE Transactions on Image Processing*, 18(6):1326–1339, 2009. (Cited page 25.)

[112] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 588–595, 2014. (Cited page 34.)

[113] Vincentius JA Verlinden, Jos N van der Geest, Albert Hofman, and M Arfan Ikram. Cognition and gait show a distinct pattern of association in the general population. *Alzheimer's & Dementia*, 10(3):328–335, 2014. (Cited page 86.)

[114] Antonio W. Vieira, Erickson R. Nascimento, Gabriel L. Oliveira, Zicheng Liu, and Mario F.M. Campos. STOP: Space-time occupancy

patterns for 3D action recognition from depth map sequences. In *Iberoamerican Congress on Pattern Recognition*, pages 252–259, Buenos Airies, Argentina, Sept. 2012. (Cited page 31.)

[115] Chunyu Wang, Yizhou Wang, and Alan L Yuille. Mining 3d key-pose-motifs for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2639–2647, 2016. (Cited pages x, 35, and 36.)

[116] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In *Computer vision–ECCV 2012*, pages 872–885. Springer, 2012. (Cited pages ix, 20, 21, and 32.)

[117] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297. IEEE, 2012. (Cited pages xi, 34, 36, 45, 52, 68, 75, 77, and 80.)

[118] Toshiya Watanabe, Susumu Shibusawa, Masaru Kamada, Tatsuhiro Yonekura, and Naohiro Ohtsuka. Design and development of lower limb chair exercise support system with depth sensor. *Transactions on Networks and Communications*, 3(4):30, 2015. (Cited page 15.)

[119] P. Wei, Y. Zhao, N. Zheng, and S. C. Zhu. Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. *IEEE International Conference on Computer Vision (ICCV)*, pages 3272–3279, 2013. (Cited pages 4, 44, and 80.)

[120] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for event and object recognition. In *2013 IEEE International Conference on Computer Vision*, pages 3272–3279. IEEE, 2013. (Cited pages ix, x, 21, 25, 36, and 37.)

[121] Ping Wei, Nanning Zheng, Yibiao Zhao, and Song-Chun Zhu. Concurrent action detection with structural prediction. In *Proceedings of*

the IEEE International Conference on Computer Vision, pages 3136–3143, 2013. (Cited page 21.)

[122] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, et al. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14–30, 2014. (Cited page 21.)

[123] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4362–4370, 2015. (Cited pages ix, 21, 22, and 23.)

[124] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. A scalable approach to activity recognition based on object use. *IEEE Conference on Computer Vision (ICCV)*, pages 1–8, 2007. (Cited page 26.)

[125] Min-Yu Wu, Tzu-Yang Chen, Kuan-Yu Chen, and Li-Chen Fu. Daily activity recognition using the informative features from skeletal and depth data. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1628–1633. IEEE, 2016. (Cited page 38.)

[126] Lu Xia and J. K. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proc. CVPR Work. on Human Activity Understanding from 3D Data*, pages 2834–2841, Portland, Oregon, USA, June 2013. (Cited page 31.)

[127] Lu Xia, Chia-Chih Chen, and JK Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27. IEEE, 2012. (Cited pages 21 and 33.)

[128] R. Xiang, J.and Liang. Motion recognition and synthesis based on

3d sparse representation. *Signal Processing*, pages 82–93, 2015. (Cited pages 4 and 25.)

[129] Zhanping Xu, Rudolf Schwarte, Horst-Guenther Heinol, Bernd Buxbaum, and Thorsten Ringbeck. Smart pixel: photonic mixer device (pmd); new system concept of a 3d-imaging camera-on-a-chip. 2005. (Cited page 17.)

[130] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 804–811, 2014. (Cited pages 4 and 25.)

[131] Xiaodong Yang and YingLi Tian. Super normal vector for activity recognition using depth sequences. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Colombus, USA, june 2014. (Cited page 32.)

[132] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proc. ACM Int. Conf. on Multimedia*, pages 1057–1060, Nara, Japan, Oct. 2012. (Cited page 30.)

[133] Hong Cheng Yang Zhao and Lu Yang. 3d sparse quantization for feature learning in action recognition. In *IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, Chengdu, China, July 2015. (Cited page 31.)

[134] B. Yao and S. C. Zhu. Learning deformable action templates from cluttered videos. *IEEE International Conference on Computer Vision (ICCV)*, pages 1507–1514, 2009. (Cited page 27.)

[135] B. Z. Yao, B. X. Nie, Z. Liu, and S. C. Zhu. Animated pose templates for modelling and detecting human actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(3):436–452, 2014. (Cited page 27.)

[136] Bangpeng Yao and Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *Com-*

puter Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 9–16. IEEE, 2010. (Cited pages ix, 26, and 27.)

[137] Bangpeng Yao and Li Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1691–1703, 2012. (Cited pages ix and 27.)

[138] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall. A survey on human motion analysis from depth data. *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pages 149–187, 2013. (Cited pages 4 and 44.)

[139] Gang Yu, Zicheng Liu, and Junsong Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision*, pages 50–65. Springer, 2014. (Cited pages x, 20, 21, 34, 35, 61, 70, 71, 72, 77, 78, and 80.)

[140] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. IEEE, 2012. (Cited page 21.)

[141] Hao Zhang and Lynne E Parker. Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):541–555, 2016. (Cited page 38.)

[142] Jing Zhang, Wanqing Li, Philip O Ogunbona, Pichao Wang, and Chang Tang. Rgb-d-based action recognition datasets: A survey. *Pattern Recognition*, 60:86–105, 2016. (Cited page 20.)

[143] Yang Zhou, Bingbing Ni, Richang Hong, Meng Wang, and Qi Tian. Interaction part mining: A mid-level approach for fine-grained action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3323–3331, 2015. (Cited pages ix and 28.)

[144] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. *European Conference on Computer Vision (ECCV)*, pages 408–424, 2014. (Cited page 26.)

[145] M. Ziaeefard and R. Bergevin. Semantic human activity recognition: a literature review. *Pattern Recognition*, 48(8):2329–2345, 2015. (Cited page 25.)

Appendix

We need to quantify shape deformations of curves involving bending, stretching and shrinking. Consider the two curves. Let us fix the parametrization of one of them to be arc-length. That is we are going to traverse that curve with speed equal to one. In order to better match the curve with the down one, one should know at what rate we are going to move along the second curve so that points reached at the same time on two curves are as close as possible under some geometric criterion. In other words, peaks and valleys should be reached at the same time. An elastic metric is the measure of that shrinking. This is termed an *elastic matching* [102].

Elastic metric

Let B the set of all parametrized curves in $\mathbb{R}^n$. $\beta : I \to \mathbb{R}^n$. β is supposed to be continuous and $\dot{\beta}(t)$ exists almost everywhere.

Let β an element of B which derivative never vanishes: $\dot{\beta}(t) \neq 0, \forall t$. $\dot{\beta}$ can be seen as: $\dot{\beta}(t) = \exp(\phi(t))v(t)$, where ϕ represent the log-speed and $v(t)$ represent the direction vector as: $\phi(t) = log(\|\dot{\beta}(t)\|)$ and $v(t) = log(\|\dot{\beta}(t)\|)$. Clearly $v(t)$ and ϕ completely specify $\dot{\beta}$ and the curve is seen as element in $\Phi \times \Upsilon$, where $\Phi = \{\phi : [0,1] \to \mathbb{R}\}$ and $\Upsilon = \{v : [0,1] \to \mathbb{S}^{n-1}\}$. Intuitively, ϕ tells us the (log of the) speed of traversal of the curve, while v tells us direction of the curve at each time t. In order to quantify the magnitudes of perturbations of β, a metric on $\Phi \times \Upsilon$ should be putted. First we note that the tangent space of $\Phi \times \Upsilon$ at any point (ϕ, v) is given by

$$T_{(\phi,v)}(\Phi, \Upsilon) = \{(u,v) : u \in T_\phi\Phi, v \in T_v\mathbb{S}^{n-1}\} \qquad (7.1)$$

with $T_\phi\Phi = \Phi$. as it is a linear space. Υ is a hypersphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^n)$ and its tangent space is given by:

$$T_v\Upsilon = \{f, f : [0,1] \to \mathbb{R}^n \ , \forall t, < f(t), v(t) >= 0\} \qquad (7.2)$$

Suppose (u^1, f^1) and (u^2, f^2) are both elements of $T_{(\phi,v)}(\Phi, \Upsilon)$ and let a et b be positive numbers.

Definition [102]*(**Elastic Metric**)*. For every point $(\phi, v) \in (\Phi \times \Upsilon)$, define an inner product on the tangent space $T_{(\phi,v)}(\Phi, \Upsilon)$ as:

$$< (u^1, f^1), (u^2, f^2) >= a^2 \int_0^1 u^1(t)u^2(t) \exp(\phi(t))\, \mathrm{d}t + b^2 \int_0^1 < f^1(t), f^2(t) > \exp(\phi(t))\, \mathrm{d}t \tag{7.3}$$

Note that $<, >$ in the second integral on the right denotes the standard dot product in $\mathbb{R}^n$. This elastic metric has the interpretation that the first integral measures the amount of *stretching*, since u^1 and u^2 are variations of the log speed ϕ of the curve, while the second integral measures the amount of *bending*, since f^1 and f^2 are variations of the direction v of the curve [102]. Therefore, the choice of weights a and b determines relative penalty on bending and stretching and a family of elastic metric is formed.The use of this metric to compare curves is motivated by the fact that the groups $SO(n)$ and Γ both act by isometries. Let $O \in SO(n)$ acts on a curve β by $(O, \beta)(t) = O\beta(t)$ and $\gamma \in \Gamma$ acts on β by $(\gamma, \beta)(t)$. $O \in SO(n)$ acts on (ϕ, v) by $(O, (\phi, v)) = (\phi, Ov)$ and $\gamma \in \Gamma$ acts on (ϕ, v) by $(\gamma, (\phi, v)) = (\phi \circ \gamma + ln \circ \dot{\gamma}, v \circ \gamma)$. $O \in SO(3)$ acts by the restriction of a linear transformation on the tangent space of $\Phi \times \Upsilon$: $(O, (u, f)) = (u, Of)$, where $(u, f) \in T_{(\phi,v)}(\Phi, \Upsilon)$ and $(u, Of) \in T_{(\phi,Ov)}(\Phi, \Upsilon)$.

The action of γ given in the above formula is affine linear, because of the term $ln \circ \gamma$. This make its action on the tangent space the same, but without this additive term: $(\gamma, (u, f)) = (u \circ \gamma, v \circ \gamma)$, where $(u, f) \in T_{(\phi,v)}(\Phi, \Upsilon)$ and $(u \circ \gamma, v \circ \gamma) \in T_{(\gamma,(\phi,v))}(\Phi, \Upsilon)$. Combining the action of $SO(3)$ and Γ with the inner product presented in equation 7.3 on (Φ, Υ), it is easy to verify that these actions are by isometries, *ie*,
$< (O, (u_1, f_1)), (O, (u_2, f_2)) >_{(O,(\phi,v))} =< (u_1, f1), (u_2, f_2) >_{(\phi,v)}$
$< (\gamma, (u_1, f_1)), (\gamma, (u_2, f_2)) >_{(\gamma,(\phi,v))} =< (u_1, f1), (u_2, f_2) >_{(\phi,v)}$

We note that in-depending of the values of a and b, both the groups $SO(3)$ and Γ act by isometries. An important question is: Is there some particular choice of weights a and b to make calculus easier? We propose to use SRV representation already used in previous chapter for its simplicity of calculus. In this chapter the SRV representation finds its potential for elastically match curves.

Square Root Velocity representation SRV

In term of (ϕ, v), SRV is given by $q(t) = \exp(\frac{1}{2}\phi(t))v(t)$. The tangent vector to $\mathbb{L}^2(I, \mathbb{R}^n)$ at q is given by a simple derivation calculus as: $h = \frac{1}{2}\exp(\frac{1}{2}\phi)uv + \exp(\frac{1}{2}\phi)f$. Let (u_1, f_1) and (u_2, f_2) denote two elements of $T_{(\phi,v)}(\Phi, \Upsilon)$, and let h_1 and h_2 denote the corresponding tangent vectors to $\mathbb{L}^2(I, \mathbb{R}^n)$ at q. The $\mathbb{L}^2$ inner product of h_1 and h_2 is given by:

$$< h_1, h_2 >= \int_0^1 < \frac{1}{2}\exp(\frac{1}{2}\phi)u_1 v + \exp(\frac{1}{2}v)f_1, \frac{1}{2}\exp(\frac{1}{2}phi)u_2 v + \exp(\frac{1}{2}v)f_2 > \mathrm{d}t \tag{7.4}$$

$$< h_1, h_2 >= \int_0^1 (\frac{1}{4}\exp(\phi)u_1 u_2 + \exp(\phi) < f_1, f_2 >) \,\mathrm{d}t \tag{7.5}$$

In this computation, $v(t)$ is an element of the unit sphere hence the fact $< v(t), v(t) >= 1$ was used to reduce the formula. It was used also that $< v, f_i(t) >= 0$ since each $f_i(t)$ is a tangent vector to the unit sphere at $v(t)$.

This expression illustrates a particular elastic metric: for $a = \frac{1}{2}$ and $b = 1$. Therefore, the $\mathbb{L}^2$ metric on the shape of SRV representations corresponds to the elastic metric on $\Phi \times \Upsilon$ and this makes the calculus simpler. Actually, expressed in terms of SRV, the $\mathbb{L}^2$ metric does not depend on the point at which these tangent vectors are defined. Finally, the inner product is simply given by:

$$< h_1, h_2 >= \int_0^1 < h_1(t), h_2(t) > \mathrm{d}t \tag{7.6}$$

In term of β, the SRV map is defined as: $SRV : B \rightarrow \mathbb{L}^2(I, \mathbb{R}^n)$

$$q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}} . \tag{7.7}$$

if $\dot{\beta}(t) \neq 0$ and 0 otherwise.

Riemannian elastic metric on open curves

Let $\beta : I = [0,1] \to \mathbb{R}^n$, represent a curve. To analyze the shape of β, we shall represent it mathematically using the *square-root velocity function* (SRVF), denoted by $q(t)$, according to:

$$q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}} \,. \tag{7.8}$$

$q(t)$ is a special function of β that we already used in previous chapter for its simplicity of calculus. Actually, the classical elastic metric for comparing shapes of curves becomes the $\mathbb{L}^2$-metric under the SRVF representation. This point is very important as it simplifies the calculus of elastic metric to the well-known calculus of functional analysis under the $\mathbb{L}^2$-metric. Hence, the SRV representation finds its potential for its ability for elastic matching. Actually, under $\mathbb{L}^2$-metric, the re-parametrization group acts by isometry on the manifold of q function (or SRV representation). This is not valid in the case of β. More formally, let β_1 and β_2 represent two curves and $\Gamma = \{\gamma : [0,1] \to [0,1], \gamma \text{ is a diffeomorphism }\}$ the set of all re-parametrizations.

$$\|\beta_1 - \beta_2\| \neq \|\beta_1 \circ \gamma - \beta_2 \circ \gamma\|. \tag{7.9}$$

The use of SRV representation allows the re-parametrization group to act by isometry on the manifold of SRV representations. This point is very important as the curve matching could be done after re-parametrization. The change of parametrization before the matching is able to reduce the effect of stretching and/or stretching of the curve.

We define the set (Pres-shape space):

$$\mathcal{C} = \{q : I \to \mathbb{R}^n, \|q\| = 1\} \ \subset \ \mathbb{L}^2(I, \mathbb{R}^n) \,. \tag{7.10}$$

The closure condition described in previous chapter is no more necessary in this case. This simplifies the calculus, with the $\mathbb{L}^2$ metric on its tangent spaces, $\mathcal{C}$ becomes a Riemannian manifold. In particular, since the elements of $\mathcal{C}$ have a unit $\mathbb{L}^2$ norm, $\mathcal{C}$ is a hypersphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^n)$. In order to compare the shapes of two curves, we can compute the distance between them in $\mathcal{C}$ under the chosen metric. This distance is defined to be the length of a geodesic connecting the two points in $\mathcal{C}$. Since $\mathcal{C}$ is a sphere, the geodesic length between any two points $q_1, q_2 \in \mathcal{C}$ is given by:

$$d_c(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle) \ , \tag{7.11}$$

and the geodesic path $\psi : [0, 1] \to \mathcal{C}$, is given by:

$$\psi(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta) q_1 + \sin(\theta\tau) q_2\right) \ ,$$

where $\theta = d_c(q_1, q_2)$. Figure **??** illustrates the space $\mathcal{C}$ and geodesic path between two elements of that space. As illustrated in Figure **??**, the space of all curves is a sphere in Hilbert space. Thus, the geodesic on the space of curves is the arc of the great circle connecting the two curves seen as elements of this sphere.

As we did in last chapter, we define the equivalent class containing q as:

$$[q] = \{\sqrt{\dot{\gamma}(t)} O q(\gamma(t)) | O \in SO(3), \ \ \gamma \in \Gamma\} \ ,$$

to be equivalent from the perspective of shape analysis. The set of such equivalence classes, denoted by $\mathcal{S} \doteq \mathcal{C}/(SO(3) \times \Gamma)$ is called the *shape space* of open curves in $\mathbb{R}^n$. $\mathcal{S}$ inherits a Riemannian metric from the larger space $\mathcal{C}$ due to the quotient structure [102].

Thanks to SRV representation, the groups $\Gamma \times SO(3)$ act by isometries. This is a necessary condition to let the quotient space $\mathcal{S}$ inherit the metric from the pre-shape space $\mathcal{C}$.

To obtain geodesics and geodesic distances between elements of $\mathcal{S}$, one needs to solve the optimization problem:

$$(O^*, \gamma^*) = \arg .min_{(O,\gamma) \in SO(3) \times \Gamma} d_c(q_1, \sqrt{\dot{\gamma}} O (q_2 \circ \gamma)) \ .$$

For a fixed O in $SO(3)$, the optimization over Γ is done using Dynamic Programming. More description of this optimization method is given in the next section. Similarly, for a fixed $\gamma \in \Gamma$, the optimization over $SO(3)$ is performed using Singular Value Decomposition method.

By iterating between these two, we can reach a solution for the joint optimization problem. Let $q_2^*(t) = \sqrt{\dot{\gamma^*(t)}}O^*q_2(\gamma^*(t)))$ be the optimal element of $[q_2]$, associated with the optimal rotation O^* and re-parameterization γ^* of the second curve, then

$$d_s([q_1],[q_2]) \doteq d_c(q_1,q_2^*)\ , \tag{7.12}$$

and the shortest geodesic between $[q_1]$ and $[q_2]$ in $\mathcal{S}$ is given by:

$$\psi(\tau) = \frac{1}{\sin(\theta)}\left(\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2^*\right)$$

,
where θ is now $d_s([q_1],[q_2])$.

Optimal Re-parametrization for curve matching

For the current rotation $O \in SO(n)$, let $\hat{q_2} = Oq_2$ and define a cost function $H : \Gamma \to \mathbb{R}_{>=0}$ by:

$$H(\gamma) = \int_0^1 \|q_1(t) - \sqrt{\dot{\gamma(t)}}\hat{q_2}(\gamma(t)\|^2\,\mathrm{d}t \tag{7.13}$$

In order to find the optimal re-parametrization, we need to find a minimum of H in Γ. Several methods allow to do that like dynamic programming [91]. The cost function H is additive over the path $(t,\gamma(t))$.

Dynamic programming

Dynamic Programming algorithm (DP) [91], was first introduced in 1962 by Bellman and Dreyfus to solve matching problem. The key idea behind dynamic programming is quite simple. In general, to solve a given problem, we need to solve different parts of the problem (sub-problems), then combine the solutions of the sub-problems to reach an overall solution. Often, many of these sub-problems are really the same. The dynamic programming approach seeks to solve each sub-problem only once, thus saving a lot of computation. This is especially useful when the number of repeating sub-problems is exponentially large. This method is based on minimization of a certain type of cost function. This cost function has to be additive in time t, which the case in the cost function presented in equation 7.13. γ is seen as a graph from $(0,0)$ to $(1,1)$ in $\mathbb{R}^2$ such that the slop of this graph is always strictly between 0 and 90 degrees. To decompose the large problem into several sub-problems, define a partial cost function:

$$E(s,t;\gamma) = \int_s^t \|q_1(t) - \sqrt{\dot{\gamma(\tau)}}\hat{q_2}(\gamma(\tau))\|^2 \, \mathrm{d}\tau \tag{7.14}$$

so the original cost function defined in 7.13 is simply $E(0,1;\gamma)$. The computer implementation details are provided in the next section.

Computer implementation

We remind that our goal is to find an optimal path from $(0,0)$ to $(1,1)$ in $\mathbb{R}^2$, corresponding to $(t,\gamma(t))$ that minimizes the cost function presented in equation 7.13. In order to use a numerical approach, the domain $[0,1] \times [0,1]$ is replaced with a finite grid and we restrict over search to that grid. The grid $G_n \times G_n$ is formed by uniform partition of G_n as $G_n = \{1/n, 2/n, ..., (n-1)/n, 1\}$. The search will be done over the set of all restrictions of γ to this grid. The total cost associated with the path is the sum of the costs associated with its linear segments. On an $n \times n$ grid there are only a finite number of paths, even less when we impose the slope constraint. Actually the path is never vertical or horizontal. However, this number of paths grows exponentially with n and we can not possibly search over all possible paths in an exhaustive fashion. Instead, the DP finds the optimal path in $O(n^2)$ time.

Denote a point on the grid $(i/n, j/n)$ by (i,j). Certain nodes are not allowed to go to (i,j) due to the slope constraint. Denote by N_{ij} the set of nodes that are allowed to go to (i,j). For instance $N(i,j) = \{(k,l)/0 \leq k < i; l < j \leq n\}$ is a valid set. Let $L(k,l;i,j)$ denote a straight line joining the nodes (k,l) and (i,j); for $(k,l) \in N_{ij}$ this is a line with slope strictly between 0 and 90 degrees. This sets up the iterative optimization problem:

$$(\hat{k},\hat{l}) = \arg.min_{(k,l)\in N_{ij}} L(k,l;i,j)\ . \tag{7.15}$$

with E as defined in equation 7.14. Define the minimum energy of reaching the point (i,j), in an iterative fashion as:

Algorithm 2: Dynamic Programming Algorithm.

```
$E = zeros(n,n);$
$E(1,:) = 1;$
$E(:,1) = 1;$
$E(1,1) = 0;$
for $i \leftarrow 2$ to $n$ do
    for $j \leftarrow 2$ to $n$ do
        for $Num \leftarrow 1$ to $size\ (Nbrs, 1)$ do
            $k = i - Nbrs(Num,1);$
            $l = j - Nbrs(Num,2);$
            if $(k > 0\&l > 0)$ then
                $H_c(Num) =$
                 $H(k,l) + FunctionE(q_1, \sqrt{\dot{\gamma(\tau)}}q_2, k/n, l/n, j/n);$
            else
                $H_c(Num) = C;$
            $H(i,j) = min(H_c);$
```

In this algorithm, C is a large positive number, $FunctionE$ is a subroutine that computes $E(\hat{k}/n, \hat{l}/n; L(\hat{k},\hat{l};i,j))$ the partial cost function defined in equation 7.14, and $Nbrs$ is a list of sites used to define N_{ij}. Typically, $Nbrs$ is a two-column matrix of type $\{(1,1);(1,2);(1,3);(2,3);(3,1);(3,2);...,\}$ depending on the number of preceding neighbors included in the implementation.