Vision-based Human Activity Analysis

Bangli Liu

School of Computing

University of Portsmouth

A thesis submitted for the degree of

Doctor of Philosophy

September 2018

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

This thesis is dedicated to my parents, and my husband Haibin Cai, with love.

Acknowledgements

During my role as a PhD student, I received helps from many people. I would like to greatly thank my supervisor, Prof. Honghai Liu, who gives me this wonderful opportunity and constant support over the past three years. I am also very grateful to Dr. Zhaojie Ju, Dr. Xiaolong Zhou, Dr. Hui Yu who offers me help on my research studies.

I would like to thank the Intelligent Systems and Biomedical Robotics Group and laboratory for the research atmosphere and devices. Thanks to Dr. Yinfeng Fang, Mr. Dalin Zhou, Mr. Dongxu Gao, Mr. Yiming Wang, Dr. Shu Zhang, Dr. Han Liu, Ms. Kairu Li, Mr. Disi Chen, Mr Song Li, Ms Xiaohua Cui for sharing their knowledge. We have good moments together in Portsmouth.

I am grateful to my parents for their unconditional love throughout all my studies. Special thanks to Haibin Cai, who stands by me all the moments.

Abstract

Human activity recognition has been an active research topic for decades due to its potential applications in video surveillance, human-robot interaction, elderly care, and entertainment. Although significant progress has been made recently with the emergency of RGB-D sensors, it still remains a great challenge in applying it to practical scenarios. The main contribution of this thesis is a novel human activity framework including four algorithms, namely, Geometry property and Bag of Semantic moving Words (GBSW) for human action recognition, Spatial Relation and temporal Moving Similarity (SRMS) for human interaction recognition, Skeleton Motion Distribution model (SMD) for human action detection, and Multi-stage Soft Regression (MSR) framework for online human activity recognition.

Firstly, targeting at traditional human action recognition problem where the action sequences are manually pre-segmented, a spatio-temporal feature descriptor GBSW which aggregates a bag of semantic moving words (BSW) with the geometric feature (G) is proposed to effectively represent human actions from skeleton sequences. Experimental results have shown that GBSW can obtain superior performance over the state-of-the-art methods.

Secondly, taking advantage of the BSW feature extracted from individuals, the moving similarity between body parts is further explored to describe the mutual relationship for effective human interaction recognition. A new large RGB-D based human-human interaction dataset, namely, Online Human Interaction (OHI) Dataset is collected for the evaluation of human interaction recognition algorithms. The effectiveness of the proposed

method has been proven by the experimental results on both the public dataset and the newly collected dataset.

Thirdly, to remove the manual segmentation requirement in the traditional action recognition and achieve automatic action detection for a given video sequence, a novel SMD model is developed. Specifically, an adaptive density estimation function is built to calculate the density distribution of skeleton movements. Experimental results have demonstrate that our method outperforms the state-of-the-art methods in terms of both detection accuracy and recognition precision.

Fourthly, a MSR framework is developed for online activity recognition where the action needs to be recognized immediately for a continuously incoming video stream. The developed framework delicately assembles overlapped activity observations in all periods to improve its robustness against arbitrary activity segments. Extensive experimental results on several public available databases have demonstrated the outstanding performance of the MSR method over the state-of-the-art approaches.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Human conduct many activities for different purposes in daily life by interacting with objects, partners or robots. Automatically recognizing human actions using the data captured by vision sensors has attracted increasing attention due to its wide applications in many areas such as video surveillance, elderly care, entertainment, and human-machine interaction. According to the complexity, common human activities can be broadly grouped into the following four different categories: gestures, actions, interactions, and group activities (Aggarwal & Ryoo, 2011). The movements of a person's specific body part, such as 'stretching a leg' and 'nodding', are defined as gestures. It has the lowest complexity of the four activities and could be combined to represent a specific intention of a person. Actions are usually performed by a single person and may be composed of multiple gestures (e.g., 'waving hands', 'running'). Interactions involve two persons (e.g., 'shaking hands', 'hugging'). Group activities are the activities performed by groups that are composed of multiple persons (e.g. 'a group meeting', 'a group of people playing games'). Compared to gestures and actions, interactions and group activities are more complex due to the involvement of more subjects and the interdependence between each subject. This thesis targets at actions performed by a single person and interactions between human and human, with a special focus on developing algorithms to be applied into practical scenarios.

The task of human activity recognition is to automatically identify the human behaviors giving a video stream or a still image, as shown in Fig. 1.1. Feature extraction and model training are two important and necessary steps to obtain a final recognition result. The feature extraction procedure aims to build a most representative descriptor for the input data. To achieve a good classification, the extracted features are expected to have a large inter-class variety and a small intra-class variety. The model training procedure will use a classifier to train a specific model for action recognition. After the model is trained, the activity inference will assign class labels to action sequences.

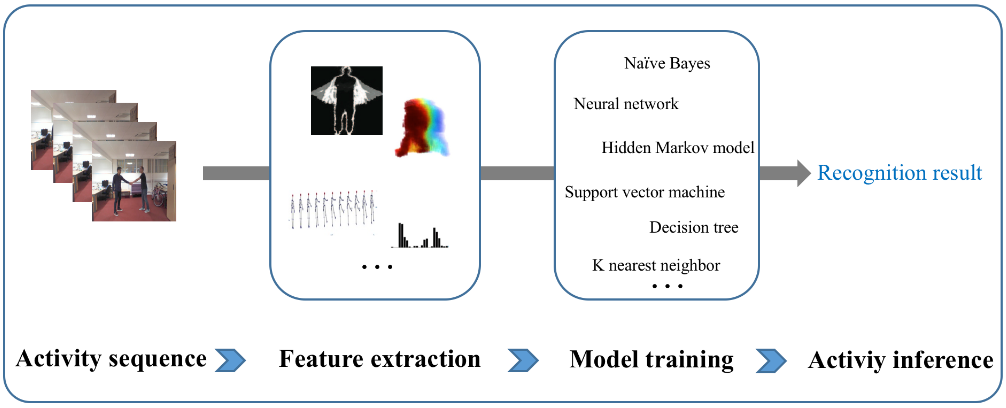


Fig. 1.1: Framework of Vision-based Human Activity Recognition.

Most of the early activity recognition research concentrates on using data captured by RGB images due to the limited sensing technology. Although this kind of data source could provide rich color and texture information about the scene, it is sensitive to illumination conditions, which will lead inconstant recognition results. Fortunately, this sensitivity can be largely alleviated by the cost-effective RGB-D sensors that provide easy access to extra depth images and 3D skeleton data. Compared to common RGB cameras, the RGB-D sensors equipped with an extra IR projector and IR camera can measure the distance information of the scene via the structured light technique. The depth information facilitates the segmentation of human bodies and the extraction of 3D body joints even in an extremely dark environment (Shotton *et al.*, 2013a). It has motivated many skeleton-based methods, depth image-based methods, and hybrid-based methods.

Traditional activity recognition algorithms focus on classifying different types of activities from pre-segmented action sequences, where each video clip only contains one complete activity. These action sequences are manually segmented after the action is finished. However, for many practical applications, the action recognition is expected to be executed with low latency while the action is being performed. For example, an alarm is expected to be triggered when dangerous behaviors happen in a public surveillance, and the assisted robot should be able to provide immediate help for the elderly people if they are going to fall down, etc. Action recognition in these realistic scenarios is referred to as online action recognition where actions need to be recognized immediately from a continuously incoming video stream. Online action recognition is challenging in that the action detection and recognition need to be conducted simultaneously in a limited time window where the action might only be partly observed. To this end, it has been one of the bottlenecks for many practical human activity recognition applications.

Based on the background, the goal of this thesis is to develop a set of new methodologies and techniques for human activity recognition in practice. The problems and challenges are summarized in Section 1.2.

1.2 Problems and Challenges

Human beings can identify an ongoing action from recorded videos easily, however, it is challenging for a computer to understand the human activity. To develop an effective and automatic human activity recognition system, the following problems and challenges need to be addressed:

1. The performance of human action recognition methods suffer from variance in individual's performing styles, body sizes and appearances, viewing conditions, and execution speeds (Shahroudy *et al.*, 2016b; Vemulapalli *et al.*, 2014; Xia *et al.*, 2012; Zanfir *et al.*, 2013). It is challenging to develop a feature descriptor which is distinctive for different activities and similar among activity sequences from the same category.

2. Compared to actions performed by a single person, human interactions are more complex in that the interdependence between each other also plays an important role. It is challenging to directly adapt existing single human action methods to interactive scenarios to achieve high recognition accuracy (Kong & Fu, 2014;

Yun *et al.*, 2012). In addition, most of the existing RGB-D based datasets target at the evaluation of human action recognition algorithms and few datasets are specifically collected for human interaction.

3. Online activity recognition aims to automatically detect and identify human activities for a continuously incoming video stream without any prior manual segmentation. Compared to traditional activity recognition, it is more challenging in that the action detection and action recognition need to be simultaneously addressed (Shan & Akella, 2014; Zhu *et al.*, 2016b). Moreover, the detected action sequences might only contain part of a complete action and it is challenging for algorithms to effectively identify activities with the limited information.

1.3 Overview of Approaches and Contributions

Considering the challenges being presented in the previous section, this thesis makes four main contributions listed as follows:

1. An effective spatio-temporal feature descriptor GBSW which aggregates the BSW with the G feature is proposed to describe human actions from skeleton sequences. The proposed BSW, which highlights the discriminative moving trend of each activity category via a kernel-based dynamic encoding algorithm, could extract the 3D moving trend of each joint in 3D space. The geometry information among skeleton joints along the whole action sequence is calculated to extract motion cues in the temporal domain. Experimental results have shown the semantic representation and the complementary effect of the aggregation of different types of features in GBSW can achieve outperforming accuracy over the state-of-the-art methods.

2. Based on the BSW feature extracted from individuals, this thesis further proposes the SRMS descriptor by modeling the spatial relation between interactive skeleton joints and the temporal moving similarity among interactive body parts for human interaction recognition. A new large RGB-D based OHI Database (Liu *et al.*, 2017a) is collected in this thesis to serve as a benchmark of human interaction recognition. Outstanding performance on the SBU Interaction

database (Yun *et al.*, 2012) and the OHI database has been achieved by the proposed method.

3. A novel skeleton motion distribution based model is proposed to detect actions in videos. By converting the unique movement characteristics of each action to its corresponding motion distribution, the occurrence frame of actions can be detected. Then, a snippet-based classifier is designed to process the observed video immediately for action classification. This classifier is performed in fragment level and can reduce the influence of false detections.

4. This thesis proposes to formulate the online action recognition as a MSR problem. Multiple score functions that measure the compatibility between a video segment and an activity label are collaboratively learned in the MSR framework. The inherent evolution of segments from adjacent performance stages is modeled by introducing a soft label strategy into the learning formulation. This soft regression has the capacity of reinforcing the capacity of distinguishing similar partial activities and the robustness to arbitrary activity fragments. Extensive experimental results on the MAD database (Huang *et al.*, 2014) and the OHI database have shown 8.9% improvement of the MSR method over the state-of-the-art approaches.

1.4 Outline of Thesis

The rest of the thesis is structured as follows:

Chapter 2 provides a comprehensive review of human action recognition, human interaction recognition, and online activity recognition, covering both hand-crafted features and learning-based features, with a special focus on data captured by RGB-D sensors. Moreover, a detailed discussion of the reviewed work in terms of adopted method types and their performance on public datasets is presented. This Chapter aims to provide readers a systemic and comprehensive background as well as achievements in the state-of-the-art algorithms.

Chapter 3 firstly introduces BSW to describe the moving trend of skeleton joints in 3D space. The moving directions are divided into several semantic moving words and the frame-level moving direction is quantified to these words via a kernel-based encoding algorithm. To extract the temporal information of skeleton joints, the G feature among joints along the whole action sequence is also calculated by the offset displacement between the initial frame and the current frame. Finally, a spatiotemporal feature descriptor GBSW which aggregates BSW with G is calculated for human action recognition. The outstanding performance of GBSW over the state-of-the-art methods is proved by the experimental results on MSR-Action3D (Li *et al.*, 2010) and Florence3D-Action (Seidenari *et al.*, 2013).

Chapter 4 proposes the moving similarity between body parts based on each subject's moving trend feature to effectively describe the mutual relationship between subjects for human interaction recognition. To facilitate the experimental evaluation, a new large RGB-D based OHI dataset is collected, which can be used as a benchmark for the evaluation of human interaction recognition. Experiments conducted on both public SBU Kinect Interaction (Yun *et al.*, 2012) dataset and newly collected OHI dataset have demonstrated the effectiveness of the proposed method.

Chapter 5 introduces a novel skeleton motion distribution based method for action detection. The occurrence frame of actions is detected depending on the change of their distribution property among each other. After detecting actions, a snippet-based classifier is designed to process the observed video immediately for action classification. Experimental results on the MAD database (Huang *et al.*, 2014) have shown its outstanding performance of accurately detecting actions in continuous videos.

Chapter 6 introduces the MSR framework to address the partial activity observation problem in online activity recognition. Multiple score functions that measure the compatibility between a video segment and an activity label are collaboratively learned in the MSR framework. The inherent evolution of segments from adjacent performance stages is modeled by introducing a soft label strategy into the learning formulation. This soft regression has the capacity of reinforcing the capacity of distinguishing similar partial activities and the robustness to arbitrary activity fragments. Extensive experimental results on the MAD database and the OHI database have demonstrated the outstanding performance of the MSR method over the state-of-the-art approaches.

Chapter 7 summaries the thesis with a discussion of the contributions and future work.

Chapter 2

Literature Review

2.1 Introduction

Human activity recognition aims to automatically analyse the human behaviors from a video stream or a still image. Most of the early research focus on activity recognition from RGB images. However, this kind of methods are sensitive to illumination and their performance is not robust due to the absence of 3D information. Fortunately, these shortcomings can be largely alleviated by the cost-effective RGB-D sensors that provide extra depth images and 3D skeleton data. As a consequence, more and more methods based on RGB-D data have been explored, which reveals a promising future direction for human activity recognition. Recently, inspired by the remarkable success of deep learning techniques in image categorization tasks (Krizhevsky *et al.*, 2012), approaches based on the deep convolutional networks which aim to learn high-level representations directly from training data have been adopted for activity recognition.

This chapter provides a detailed review of the data capturing, human action and interaction recognition, online activity recognition, and RGB-D human activity datasets. Considering that the feature extraction process varies to a large extent and plays an important role in most of the existing methods, this thesis further categorizes the existing methods according to the construction of features. Meanwhile, the overview of the classification methods is presented during the description of each method and is also discussed in Section 2.8.

The rest of this chapter is organized as follows. Section 2.2 describes Kinect sensors and 3D data. Section 2.3 and Section 2.4 review hand-crafted features based human action and interaction recognition methods respectively. Section 2.5 presents a survey on deep learning based human activity recognition. Section 2.6 provides an overview of the existing online activity recognition methods. Section 2.7 summarizes the commonly used RGB-D human activity datasets. Section 2.8 provides a comparison between hand-crafted based and deep leaning based methods along with a discussion of their performance on the most commonly used datasets.

2.2 Kinect Sensor and 3D Data

The task of vision based human activity recognition is to identify the human behaviors from the scene observed by an acquisition system. Most of the early research use only the color and texture information in 2D images provided by traditional RGB cameras. However, the sensitivity to illumination changes and subject texture variations as well as the lack of 3D information of the scene often degrade the recognition accuracy. There are mainly three categories of methods to obtain 3D data.

The first way to get the 3D information of human bodies is using a marker-based motion capture system (Mocap). In Mocap, different types of markers are placed in specific positions (such as boney regions), and the 3D position of a human body is generated by estimating the position of each marker. Mocap is able to accurately capture human pose and track it along the time resulting in high resolution data. Therefore, Mocap has been used in a wide range applications, such as animation, video games, and virtual coaches. Several motion capture datasets have been collected providing such data for human activity analysis, like the Carnegie Mellon University Motion Capture database and HDM05 (Müller *et al.*, 2007). Fig. 2.1 shows an example of motion capture systems used by Müller *et al.* (2007). The retro-reflective markers attached to the actor's body are tracked by an array of six to twelve calibrated high-resolution cameras arranged in a circle. Although the skeleton joints obtained by the Mocap system is reliable and less noisy, this type of systems require the user to wear some physical markers to acquire the 3D data, which makes it not convenient for the general public.

The second way is to reconstruct 3D information from 2D image sequences captured from multiple views (Argyriou *et al.*, 2010). The low-cost stereo camera is equipped with two or more lenses with a separate image sensor or film frame for each

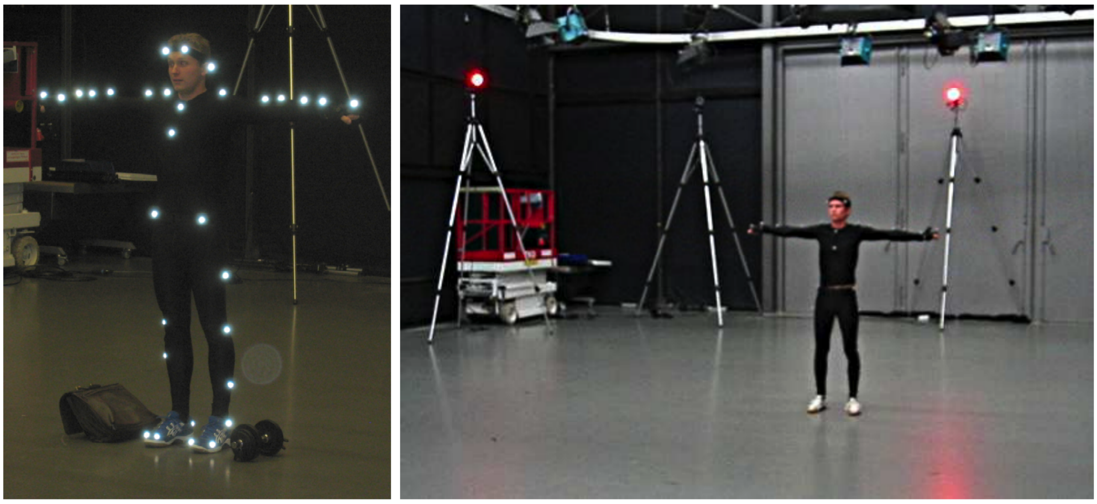
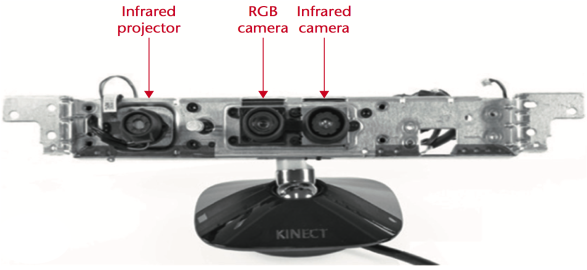


Fig. 2.1: Motion capture system used in HDM05 dataset (Müller *et al.*, 2007).

lens. This allows the camera to simulate human binocular vision and therefore gives it the ability to generate 3D images. The relative depth information to the objects could be obtained by comparing the two images. However, it is still challenging to reconstruct 3D information from stereo images due to the complexity of the geometry. In addition, the exhaustive calibration and synchronization process among multiple cameras is not desirable and limits its practical applications.

Fortunately, these shortcomings can be largely alleviated by the cost-effective Kinect sensors, such as Microsoft Kinect and ASUS Xtion, that provide easy access to the 3D structure of the scene using one camera, as shown in Fig. 2.2. The Kinect sensor is originally introduced as a game tool, which has greatly revolutionized the way people



(a)

Fig. 2.2: The structure of the Kinect sensor, including infrared (IR) projector, IR camera, and RGB camera.

play games and how they experience entertainment. It allows users to naturally interact with a computer with gestures. Compared to traditional RGB cameras, the third dimension (depth) of users provided by the Kinect sensor makes computer vision tasks such as body language understanding much easier. With its excellence and low cost, Kinect's impact has extended far beyond the gaming industry.

In addition to standard RGB images, a depth map is also provided giving each pixel the corresponding distance with respect to the sensor. Structured light and time of flight technology are two common ways to estimate the depth information. Structure light cameras (e.g. Microsoft Kinect v1) project a known pattern onto the scene and calculate its distortion to estimate the distance of points. Time of flight cameras (e.g. Micorsoft Kinect v2) emit a light signal in the scene and compute depth based on the time elapsed between the emission of a light signal and its reception with the known speed of light.

Two main software libraries namely Microsoft Kinect SDK and OpenNI SDK are developed for providing skeleton joints. Shotton *et al.* (2013a) provided an real-time effective skeleton construction algorithm based on body part distribution in 2011 (Fig. 2.3). A single depth image is classified at each pixel using a randomized decision forest classifier. Each branch in the forest is determined by a simple relation between the target pixel and various others. The pixels that are classified into the same category form the body part, whose local modes are estimated to provide high-quality proposals for 3D skeleton joints. Finally, the 3D skeleton is inferred by fitting the joint proposals with the silhouette. This algorithm is able to generate 3D human skeleton models within about 5 ms.

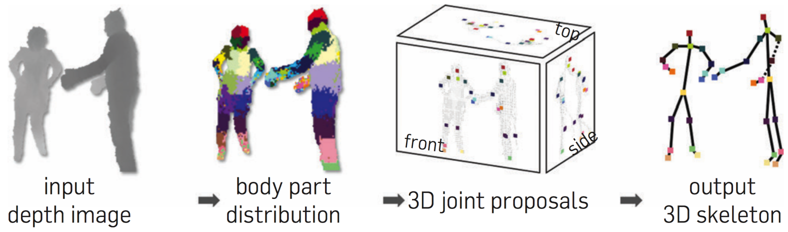


Fig. 2.3: Skeleton estimation from a single input depth image based on body part recognition (Shotton *et al.*, 2013a).

To this end, the RGB image of the scene, its corresponding depth map, and 3d skeleton information could be obtained by the Kinect sensor, as shown in Fig. 2.4. Compared to traditional RGB cameras which have high sensibility to color and light conditions, the Kinect sensor has the robustness to the change of illumination conditions. In addition, the available 3D information of the scene indeed facilitates the subtraction of objects of interest from the background. As a consequence, many methods based on RGB-D data have been explored (Han *et al.*, 2017; Presti & La Cascia, 2016; Zhang *et al.*, 2016; Zhu *et al.*, 2016a) for human activity recognition. Following sections will provide a comprehensive review of the existing human action and interaction recognition methods, including both hand-crafted based and deep learning based algorithms. Fig. 2.5 shows the categories of these methods.

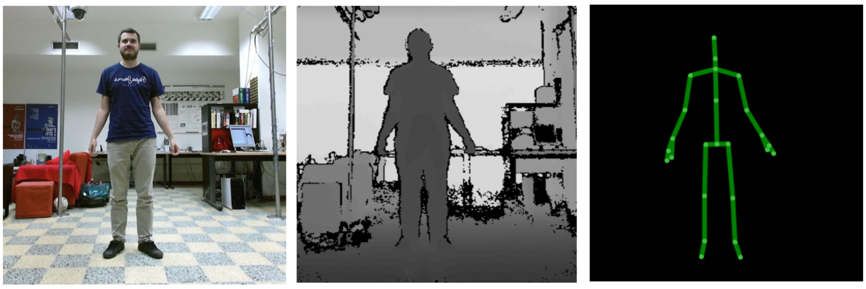


Fig. 2.4: The RGB image, depth iamge and skeleton joints obtained by the Kinect sensor.

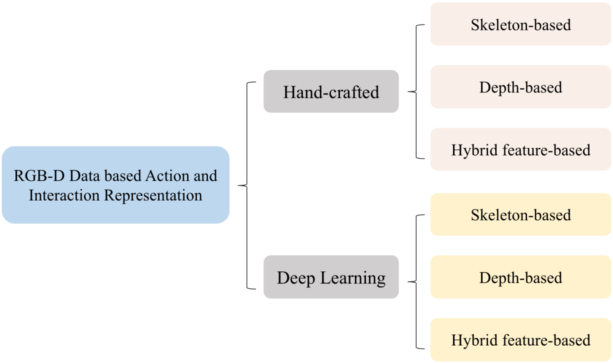


Fig. 2.5: The category of human activity representations using RGB-D data.

2.3 Hand-crafted Features based Human Action Recognition

This section mainly focuses on reviewing related works on RGB-D data based single human action recognition. Based on the data modality used, they can be classified into three categories: skeleton-based methods, depth-based methods, and hybrid feature-based methods.

2.3.1 Skeleton-based Methods

Skeleton information can be estimated using the RGB data (Toshev & Szegedy, 2014; Yang & Ramanan, 2013), wearable motion capture (Mo-Cap) sensor (Deng *et al.*, 2012) or depth data (Shotton *et al.*, 2013b; Ye *et al.*, 2011). The review of skeleton-based methods gives priority to depth data in this section. The release of RGB-D sensors such as Kinect and Xtion makes it possible to obtain 3D positions of body joints from depth images frame-by-frame (Shotton *et al.*, 2013b), encouraging a lot of recognition methods using skeleton data being proposed. The skeleton-based methods can be further divided into trajectory-based and pose-based algorithms.

Trajectory-based algorithms explore characteristics of the spatio-temporal trajectory of skeleton joints to identify various actions (Gowayyed *et al.*, 2013; Ofli *et al.*, 2014; Ohn-Bar & Trivedi, 2013; Qiao *et al.*, 2017; Zanfir *et al.*, 2013). Gowayyed *et al.* (2013) proposed a 3D trajectory descriptor, which concatenated three 2D projections of the whole skeleton sequences, to represent the movement of each joint. The final action classification was conducted based on these trajectory descriptors. In (Ohn-Bar & Trivedi, 2013), joint angles between connected pairs of body parts were chosen as motion features and then similarities between each angle with temporal evolution were used as representation of actions. A modified histogram of oriented gradients, i.e., HOGs, was utilized to capture the posture information around each joint over the whole time. Qiao *et al.* (2017) applied a local feature representation named trajectorylet, which constrained the dynamic characteristic of actions from the entire sequence to a short temporal range, to capture ample static and dynamic information of actions. Compared to extracting dynamic characteristic of actions from the entire sequence, more specific dynamic information within a short temporal range was captured in this trajectorylet description. In (Ofli *et al.*, 2014), the most informative joints of performing each action were firstly captured within an instant time according to the mean or variance of joint angle trajectories. Although this local trajectory representation could remove the inactive frames and emphasize the distinctive sub-sequences, it is challenging to extract salient trajectories from the noisy skeleton sequences. Guo *et al.* (2018) encoded motion trajectories of body parts to a gradient variation based sparse histogram and applied a support vector machine (SVM) with chi-square kernel for action recognition.

Compared to the trajectory-based approaches, pose-based approaches focus more on key poses characterized by the skeleton point distribution or its surrounding body parts. Features such as joint locations, joint angles, and 3D geometric relationships between body parts are often directly employed as advantageous representations of activities (Lillo *et al.*, 2017; Pazhoumand-Dar *et al.*, 2015; Theodorakopoulos *et al.*, 2014; Xia *et al.*, 2012). In (Xia *et al.*, 2012), the histogram of 3D joint locations, i.e., HOJ3D, in a modified spherical coordinate centering at hip center, was proposed to describe human postures which were then clustered into K clusters using K-means. These K clusters representing the prototypical poses of actions were considered as observation symbols in a discrete Hidden Markov Model (HMM) to explain the temporal evolution. Pazhoumand-Dar *et al.* (2015) applied joint angles to depict body poses and simultaneously used the relative motions between joints to describe their relationships in the time domain. To reliably identify actions from noisy skeletal data sequences, they formulated a classification function based on an extended formulation of the longest common subsequence algorithm. Instead of using the movement of all skeleton joints, Eweiwi *et al.* (2015) focused on mining discriminative joints with apparent motion property. Discriminative joints were determined by partial least squares, whose location information, velocity, and the movement normals were encoded as poses during a short video period. Chaaraoui *et al.* (2014) used Dynamic Time Warping, i.e., DTW, to calculate the matching between action sequences for action recognition, where an evolutionary algorithm was proposed to select the optimal set of skeleton joints to form sequences of key poses for each action. Vemulapalli *et al.* (2014) made use of the rotations and translations among body parts to model their relative 3D geometry relation, with which human motion was encoded as curves in Lie group. This method is able to reveal the concurrence of body parts, whereas the isolation of body parts remains difficult when there is overlapped areas among body parts.

2.3.2 Depth-based Methods

The depth images, which store the Euclidean distance between the sensor and points in the scene, make it easy to extract human bodies from the cluttered background. Furthermore, depth images are invariant to the change of lighting conditions thus stable and rich information could be provided for describing human shape or motion. Some researchers (Bulbul *et al.*, 2015; Chen *et al.*, 2013; Li *et al.*, 2010; Wang *et al.*, 2015d; Yang *et al.*, 2012a) proposed to project the 3D depth information onto three 2D orthogonal planes corresponding to the front, side, and top view for feature extraction, as shown in Fig. 2.6. Li *et al.* (2010) extracted 3D representative points of the body silhouette from these planes to model postures for recognition. The bi-gram maximum likelihood decoding algorithm was employed to reduce the computational complexity. In (Yang *et al.*, 2012a), Depth Motion Maps, i.e., DMMs, were generated by stacking depth maps over the whole sequence and then HOGs were computed to characterize human motions. The concatenation of HOGs was the input of a linear SVM classifier for final action classification. To reduce the computation cost caused by the computation of HOGs in (Yang *et al.*, 2012a), Chen *et al.* (2013) used the concatenation of DMMs from three viewpoints as the final representation. Human body

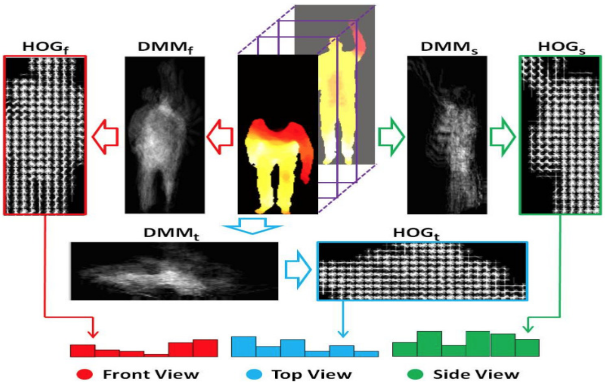


Fig. 2.6: Method proposed in (Yang *et al.*, 2012a).

shape was reflected in detail from different viewpoints in these DMMs. Bulbul *et al.* (2015) improved DMMs by implementing the contourlet transform with a multi-scale and multi-directional analysis to enhance the shape characteristic of DMMs.

The surface normal vectors calculated using a group of 3D points can be used to describe the shape and motion information (Oreifej & Liu, 2013; Slama *et al.*, 2014; Yang & Tian, 2014b). Shape and motion information were jointly extracted in a 4D space formed by extra considering the depth and time domain. Oreifej & Liu (2013) proposed to divide the depth sequences into many spatio-temporal cells and surface normals in each cell were counted to compute the Histogram of Oriented 4D, i.e., HON4D. The normal variable from human body surface depicts the change of human shape and motion. Similarly, super normal vector (Yang & Tian, 2014b), i.e., SNV, was calculated by grouping local hypersurface normals to create the low-level polynormal, which further preserves the correlation among local normals in the polynormal and achieved 2.3% improvement compared to HON4D. Slama *et al.* (2014) modeled sequence features as subspaces lying on Grassmann manifold, where the geometric and dynamic information of human body were computed. The action classification benefits from the geometric structure of Grassmann manifold. Principal component analysis, whose computational complexity was $O(min(m^3, m^2n))$ with m being the feature dimension and r being the number of training samples, was applied to reduce the dimension of features.

Alternatively, some researchers propsed to segment the depth data to interest areas (the most active parts), from which compact features were extracted for action recognition. In addtion, features from the interest areas can describe local parts of the actions thus are robust to occlusions or clutter. For example, Wang *et al.* (2012a) constructed random occupancy patterns feature from 4D subvolumes randomly sampled in depth map sequences to gain the robustness towards occulsions. Xia & Aggarwal (2013) utilized the depth cuboid similarity to depict the local feature around the spatio-temporal interest points extracted from depth videos. In (Liu & Liu, 2016), the spatial relationship among selected joints with discriminative shape and movement was used to build the depth context descriptor for final action recognition. Compared to the methods extracting features direct from images, these approaches require extra computational cost to detect interest regions through the whole depth sequence before feature extraction.

2.3.3 Hybrid Feature-based Methods

The association of multi-modal data such as skeleton data, color, and depth images can improve the recognition performance (Althloothi *et al.*, 2014; Raman & Maybank, 2016; Wang *et al.*, 2012b, 2014; Yang & Tian, 2014a; Zhu *et al.*, 2013). For example, the depth or RGB information can reflect the appearance or texture information surrounding the skeleton joints while the movement of joints can describe the motion information.

The majority of hybrid features tend to extract the corresponding depth information around skeleton joints (Raman & Maybank, 2016; Shahroudy *et al.*, 2016b; Wang *et al.*, 2012b, 2014), or combine the features from joints and depth images directly (Althloothi *et al.*, 2014; Jalal *et al.*, 2017; Ji *et al.*, 2018). Wang *et al.* (2012b, 2014) proposed the local occupancy pattern, i.e., LOP, to describe the appearance around each joint. The body movement was interpreted by the relative positions of joint pairs and LOP was obtained by projecting cloud points in the spatial grid around each joint. Ji *et al.* (2018) partitioned the human body to several motion parts by embedding the skeleton data into depth sequences. Local features extracted from these motion parts were aggregated into a discriminative descriptor. The depth inforamtion of objects around joints was also associated in (Raman & Maybank, 2016) as the low-level layer of a hierarchical HMM. It has a computational complexity of $O(TK^2)$ where T is the sequence length and K is the state number. Zhu *et al.* (2013) coupled the motion depending on points of interest and spatial information using a random forests-based fusion strategy. The frame-level fusion strategy used in this method makes information from different data sources complement each other more effectively. To reduce the confusing frames and computation cost, the information from depth images was employed to determine the discriminative frames among the whole sequence in (Yang & Tian, 2014a). Yang & Tian (2014a) proposed a depth map based accumulated motion energy function to select the discriminative skeleton frames to remove noisy frames and reduce computational cost. After the calculation of EigenJoints (Fig. 2.7) which combines the difference of postures, motion, and offset information of joints, they used non-parametric Naive-Bayes-Nearest-Neighbor (Boiman *et al.*, 2008) to classify multiple actions.

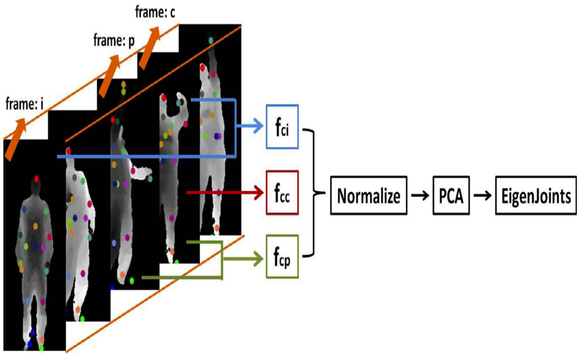


Fig. 2.7: Method proposed in (Yang & Tian, 2012).

Apart from the combination of skeleton joints and depth images, some researchers also consider RGB information (Kong & Fu, 2015a; Liu *et al.*, 2015; Sung *et al.*, 2012; Zhang & Parker, 2016). Sung *et al.* (2012) employed skeleton joints to model motion features and extracted HOGs features from regions of interest in both RGB and depth images to characterize the appearance cues. A coupled hidden conditional random fields model (Liu *et al.*, 2015) was proposed to learn the latent correlation between visual features from both RGB and depth source. In this model, the temporal context within individual modality was preserved while learning the correlation between two modalities. Alternatively, Kong & Fu (2015a) projected features from RGB and depth images into an united space and independent private spaces for action recognition, which indicated that knowledge and correlation from different sources could be shared with each other to reduce noise and improve the action recognition performance.

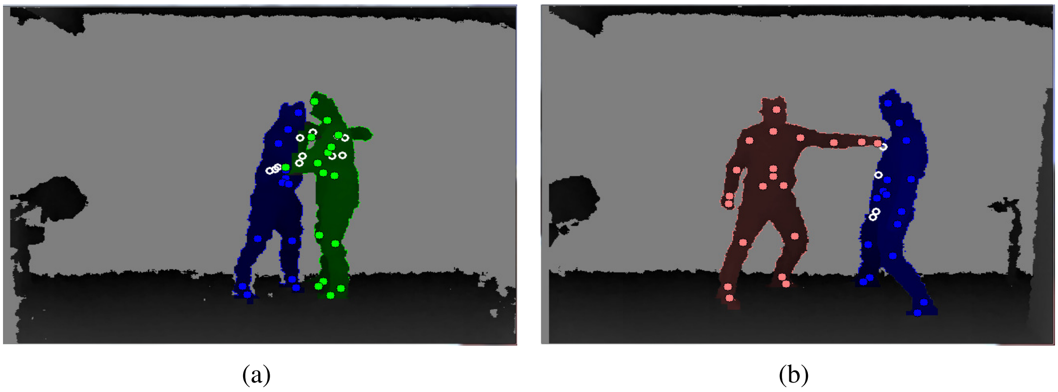
2.3.4 Summary

The hand-crafted human action recognition methods were divided into skeleton-based, depth-based, and hybrid feature-based according to the used data modality. Each type of methods has its advantages and limitations. Although depth images have the outstanding capability to describe appearance information, they might suffer from holes where depth data is missing. Skeletons are compact and straightforward to depict the motion properties of actions, however, the limitation of the skeletal feature is that it does not give information about the surrounding objects which should be considered when modeling human-object interaction. In addition, the skeleton tracking from

Kinect sensors is not very reliable when the human body is partly occluded or the subject is not in an upright position facing the sensor. The combination of features from different modalities has the potential to improve the recognition performance by overcoming the respective weakness.

2.4 Hand-crafted Features based Human Interaction Recognition

This section provides a comprehensive review of human interaction recognition using RGB-D data. They can be regarded as a type of activities where one person adapts his/her behavior according to the action of the other person. Although an interaction is a collection of atomic actions from individuals' actions, it cannot be easily isolated, especially when human bodies have overlapped area caused by physical contact or occlusion, as shown in Fig. 2.8. In these situations, there is large ambiguity of feature assignment to a unique person, which makes features used for atomic action recognition, such as interest points and trajectories, difficult to directly be applied for human interaction recognition (Kong & Fu, 2016). Compared with single person action, the feature space of human interaction has more variations in subject appearance, scale, viewpoint, interacting motion patterns, etc., due to multiple persons involved (Kong *et al.*, 2012). Moreover, diverse interacting motion patterns (the actions and reactions


(a)

(b)

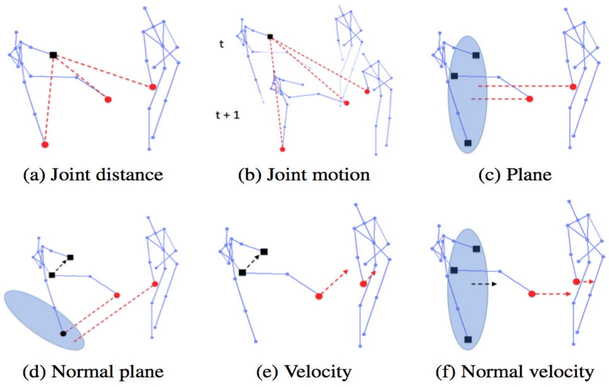
Fig. 2.8: Inaccurate estimation of skeleton joints (in white color) caused by inter-occlusion and self-occlusion during human interaction.

vary from each other) make human interaction recognition more challenging. For example, semantic performances of the defender to protect oneself vary from step back, crouch, to hit back, so this requires all possible co-occurrence relations to be extracted.

The majority of existing RGB-D data based human interaction recognition use features from skeleton sequences or combine features from different data modalities, while few approaches are based on single depth data (Gori *et al.*, 2017; Yun *et al.*, 2012). Therefore, they can be classified into two categories: skeleton-based methods and hybrid feature-based methods.

2.4.1 Skeleton-based Methods

Some human interaction recognition algorithms utilize features of joint pairs to exhibit the spatial and motion relation over the time (Huynh-The *et al.*, 2015; Yun *et al.*, 2012). Yun *et al.* (2012) extracted different features of skeleton joints, such as distance, joint movement between consecutive frames, the geometric relationship between joints and planes, and velocity features, as shown in Fig. 2.9. To reduce the irrelevant frames, interaction sequences were depicted by a bag of body-pose via multiple instance learning. Their experimental results showed that the joint features outperform velocity features by 30%. A hierarchical model was employed by (Huynh-The *et al.*, 2015) for interaction recognition, where interaction was disjointed into topics. The correlation among low-level features, topics, and activities, was exhibited in the model.



(a) Joint distance (b) Joint motion (c) Plane

(d) Normal plane (e) Velocity (f) Normal velocity

Fig. 2.9: Diverse features proposed in (Yun *et al.*, 2012).

Interactive body part pairs are distinct among different interactions, for instance, *shaking hands* can roughly be regarded as the interaction between two hands. Thus, mining the essential interactive pairs helps to remove redundant information from the inactive body parts. For example, the activities between two persons were described as the motion and spatial relations between informative body parts in (Ji *et al.*, 2014, 2015; Saha *et al.*, 2015). The contrast mining method was applied to extract the most active body part pairs for each interaction class in (Ji *et al.*, 2014). On the basis of this work, Ji *et al.* (2015) extracted intra-inter-frame features of single or interactive pairs. The learned contrastive feature distribution model provided a discriminative description for interactions. Compared to the employment of all joints in (Yun *et al.*, 2012), these mined representations are more discriminative and not computationally expensive. However, occlusion or big inter-class similarity makes the extraction of interactive body parts difficult. Wu *et al.* (2017) proposed a human interaction feature descriptor by utilizing the static, dynamic and direction properties of the skeleton data. They addressed the interaction recognition by formulating it as a sparse group lasso problem.

Some scholars transformed the interaction problem to the single person action recognition problem (Bloom *et al.*, 2014, 2016; Hu *et al.*, 2013). The interaction between players was decomposed into two single individual actions in a computer gaming environment in (Bloom *et al.*, 2016). Each player's action was trained and classified separately, and the final interaction classification was achieved through the combinition probability of two actions. Hu *et al.* (2013) (Fig. 2.10) firstly identified the most active person according to the following two rules: the person acts firstly or the

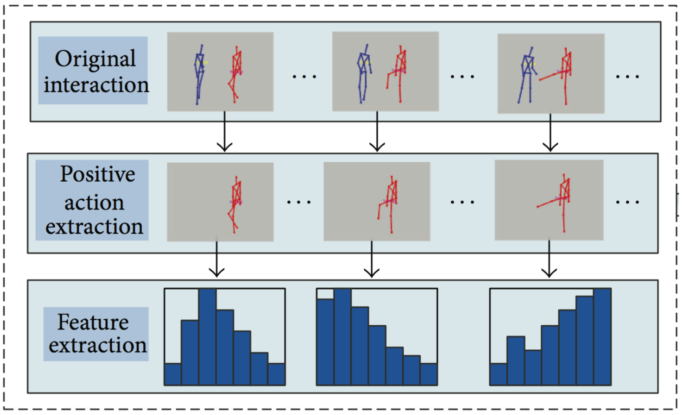


Fig. 2.10: The flowchart of positive action recognition in (Hu *et al.*, 2013).

person with greater motion at the beginning short frames.Then, the action of the active person was used for human interaction recognition. Features proposed in (Yun *et al.*, 2012) were utilized to encode the position action as a pose descriptor in (Bengalur, 2013; Hu *et al.*, 2013). Although this method is able to reduce the feature dimensions, its effectiveness will be affected when the interaction with heavy occlusion. Unlike the methods mentioned above, Coppola *et al.* (2016) utilized features from two individuals and the relationship between each other for different purposes. They treated physical proximity features learned from social interaction as prior knowledge and built a multivariate Gaussian distribution to estimate the distribution of each interaction category.

2.4.2 Hybrid Feature-based Methods

Gori *et al.* (2017) built a bounding box around the human body to restrict the area of interest because it could help to remove most of the redundant information of the different modalities. Then, a matrix called relation history image was proposed to depict the local relations, which contains Euclidean distances of joint pairs and comparison of depth value between pixels in a bounding box. Similarly, van Gemeren *et al.* (2014) explored shape and movement features for each interactive person from bounding boxes where the interaction happens and merged the information of joints with poselets to select key frames for action representation. Xia *et al.* (2015) combined the posture, motion information, and local appearance feature from both RGB and depth channel for interaction recognition. They studied interaction from a robot-centric view instead of the conditional third-person view. Appearance and motion properties were extracted directly from body parts in (Alazrai *et al.*, 2015) and (Xu *et al.*, 2015). The semantic meaning of body-part between two interacting people was described by motion direction and distance between two persons in (Alazrai *et al.*, 2015). To supplement the movement profile, local shape information was extracted from the bounding box around body parts. The final feature descriptor was formed by concatenating all these features together. In (Trabelsi *et al.*, 2017), a comprehensive feature descriptor combining the distance property of the 3D skeleton data and the dense optical feature extracted from the color and depth images was proposed to describe human interactions.

Contextual information is a vital factor for recognizing human activity, especially for those during human interaction and human-object interaction (Ni *et al.*, 2013). The relationship between activities and backgrounds, and features from a specific object with which a person interacting could support valuable context for action recognition. In (Ni *et al.*, 2013), constraints based on depth value were used to improve the accuracy of objects detection. Apart from extracting appearance features by HOGs for each subject, the spatio-temporal contextual attributes were encoded by relative distances, velocity or time order. Additionally, the depth-based environment description was considered for representing different scenes and thus made the recognition more precise. A dual assignment K-means clustering algorithm which exploits the correlation between actions and scenes was proposed in (Jones & Shao, 2014). Their experimental results showed that the performance can be improved by considering the contextual feature.

2.4.3 Summary

The multi-modal data provided by RGB-D sensors has been encouraging more researchers to investigate approaches for human interaction recognition. This section summarized both skeleton-based and hybrid feature-based methods. Currently, the interdependent relation between interactive persons was mainly represented by the distance between body parts in the proposed methods. Although the distance property is useful, it might not be sufficient to reflect the inherent relations. Thus, one direction of future research may be designing effective approaches toward the semantic interpretation of activities.

2.5 Deep Learning based Human Activity Recognition

Motivated by the great achievements of deep learning techniques in computer vision community, researchers have developed many deep learning based methods for human activity recognition. Unlike the hand-crafted methods where specific types of features need to be designed to distinguish human action and interaction, most of the deep learning based methods code human action or interaction information directly into a map and then resize the map to a fixed size for activity recognition. Therefore, this thesis does not specifically separate existing deep learning based human activity recognition methods into single human action and human interaction at this stage. Following the same taxonomy with the hand-crafted methods, the research reviewed in this section is grouped into three categories: skeleton-based, depth-based, and hybrid-feature based.

2.5.1 Skeleton-based Methods

Convolutional Neural Networks (CNN) is automatic feature extractors which could extract effective information from different spatial locations in the input image by a number of filters. Recently, inspired by its remarkable success in image categorization tasks Krizhevsky *et al.* (2012), CNN-based methods which aim to learn high-level representations directly from skeleton sequences have been adopted for action recognition. Their focus is mainly on transforming the skeleton joints positions or trajectories into images and then adapting CNN for classification.

Li *et al.* (2017) proposed to project 3D skeleton joints into three orthogonal 2D planes. Together with the 3D distance information, they constructed four joint distance maps according to a linear interpolation function and then applied AlexNet (Krizhevsky *et al.*, 2012) to classify actions. Ke *et al.* (2017) firstly transformed the coordinate of skeleton joints from Cartesian coordinates to cylindrical coordinates (Weinland *et al.*, 2006) and constructed three clips of gray images using the relative positions between the skeleton joints and four manually defined reference joints. Then, the features extracted from a pre-trained VGGNet (Simonyan & Zisserman, 2015) were fed into a multi-task learning network which consists of two fully connected layers, a rectified linear layer, and a softmax layer, for the final classification. The method dealt with the different action duration problem by coding the temporal dynamics of the skeleton sequence into respective dynamic image rows and further resizing the image to a fixed size.

Observing that the image resizing operation might introduce extra noise for the network, Liu *et al.* (2017c) proposed to directly input a skeleton image, which was constructed by indexing the skeleton joints into several tiny 5×5 images and expanded by using the spatial-temporal information, to a modified Inception-ResNet CNN architecture for action recognition, as shown in Fig. 2.11. The drawback of this method is the assumption of each action has a fixed number of skeleton sequence as input. The

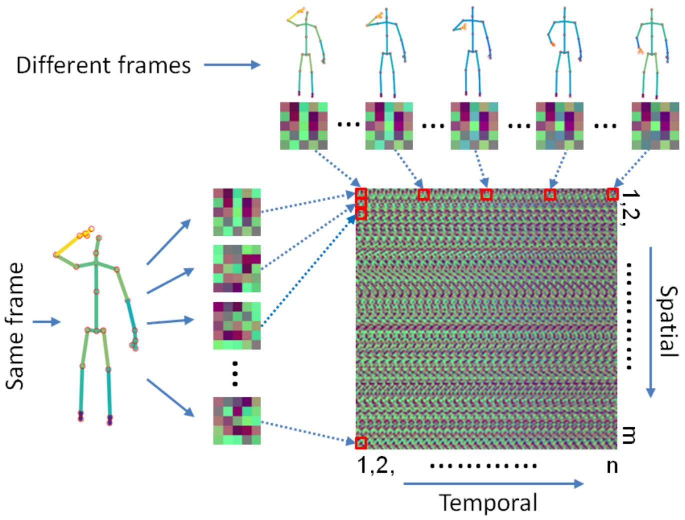


Fig. 2.11: Spatio-temporal image representation of human skeleton joints proposed in (Liu *et al.*, 2017c).

spatio-temporal information of 3D skeleton sequences was encoded into three joint trajectory maps according to three different views (i.e., front, top, and side) in (Hou *et al.*, 2018; Wang *et al.*, 2016b). A ConvNet for each trajectory map was trained and actions were classified via a late fusion of the three output scores. Similarly, Du *et al.* (2015a) proposed to code the temporal and spatial information of the skeleton sequence into image columns and rows respectively. Then a CNN with four convolution layers and two fully connected layers were used for action recognition.

Different with previous methods, Yan *et al.* (2018) proposed to employ a multi-layer graph neural networks, where the graph nodes consist of joint coordinates and estimation confidences, to automatically learn the spatio-temporal pattern of the skeleton data. The undirected spatio-temporal graph can well perseve both the connectivity of human body structure and its temporal variation in the consecutive frame, which helps to achieve over 80% recognition rate on the NTU-RGB+D dataset (Shahroudy *et al.*, 2016a). Huang *et al.* (2017) employed a neural network architecture to learn a temporally aligned Lie group representations (Vemulapalli & Chellapa, 2016) for action recognition, which demonstrated that the non-Euclidean Lie group structure can also be incorporated by the CNN structure.

As an alternative solution to CNN, Recurrent Neural Network (RNN) could effectively model the temporal information. However, due to the gradient vanishing problem, most of the RNN methods lack the ability to process long action sequences. One of the popular solutions is to employ the Long Short-Term Memory (LSTM) (Hochreiter & Urgen Schmidhuber, 1997) which solves the problem by utilizing a gating mechanism to determine the memory length of the input sequence.For example, Veeriah *et al.* (2015) proposed a differential RNN by adding a gating into LSTM to model the dynamics of salient motions. Various hand-crafted features concatenated from successive frames were fed to the proposed LSTM structure.

Du *et al.* (2015b, 2016) proposed an end-to-end hierarchical RNN which fuses the feature extracted from five human body parts for action recognition. As pointed out in (Zhang *et al.*, 2017c), the relationship between non-adjacent parts can be useful to depcit the dynamic characteristics of actions. Shahroudy *et al.* (2016a) utilized the human body structure to build a part-aware LSTM. By concatenating part-based memory cells, the non-adjacent parts relations learned from the 3d skeleton sequence help to describe the dynamics of actions and thus improve the recognition performance.

Mahasseni & Todorovic (2016) employed the regularized LSTM on top of a deep CNN for RGB video based action recognition. Assuming extra 3D skeleton data can complement the lost information in the video, they proposed to regularize the network by using the 3D skeleton sequence from a few actions. Zhu *et al.* (2016c) fed the skeleton joints to a deep LSTM network with mixed-norm regularization term to learn co-occurrence features for action recognition. They further applied an internal dropout method to the LSTM neurons in the last LSTM layer to learn complex motion dynamics, as shown in Fig. 2.12. Zhang *et al.* (2017c) explored various geometric relational

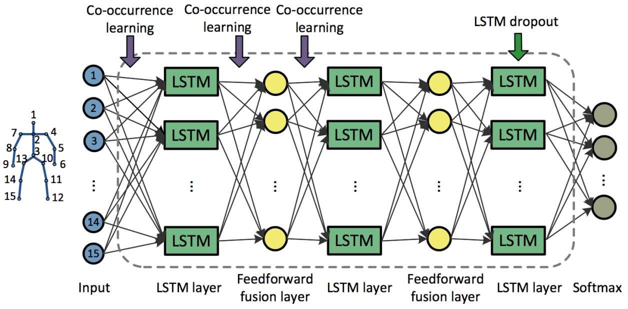


Fig. 2.12: Deep LSTM network (Zhu *et al.*, 2016c): three LSTM layere and two feed-forward layers.

features among all joints and used a stacked three layers LSTM for action recognition.

Observing the lost information in the transnational pre-processing procedure, which transforms and rotates 3D skeleton joints to the person-centric coordinate system, Zhang *et al.* (2017b) proposed a view adaptive RNN with LSTM structure to deal with the viewpoint variations. To reduce the noise of the irrelevant joints and improve the performance, Liu *et al.* (2018) developed a global context-aware attention LSTM to selectively pay attention to informative joints in each frame with the help of the global memory cell. The attention ability was further improved by using a recurrent attention mechanism. CNN and LSTM were jointly utilized for action recognition in (Núñez *et al.*, 2018), where the output of CNN was served as the input of LSTM. A two-stage training procedure was further developed, which firstly trains the CNN and then trains the combined network.

Unlike the previous RNN based methods where only the temporal domain of the skeletons are modeled, Liu *et al.* (2016b) proposed a tree-structure based traversal method to handle the spatial adjacency graph of the body joints. A trust gate was also proposed to remove noisy joints and deal with the occlusion in the 3D skeleton data. Similarly, Song *et al.* (2017) proposed to add joint-selection gates in the spatial attention model and frame-selection gazes in the temporal model for action recognition. They further proposed a joint training strategy to train the network in an end-to-end way. Wang & Wang (2017) proposed a two-stream RNN architecture which jointly models the spatial articulated property and the temporal dynamic of skeletons. The additional spatial RNN models the spatial dependency of joints by considering human body kinematics. However, for human interaction, they conducted action recognition on each person and used the average score as the final recognition result. The performance on human interaction can be improved by further considering the relationship between interactive persons.

2.5.2 Depth-based Methods

Depth image sequences contain rich motion features, however, they are not suitable to be the input of the most existing CNN models which are specifically designed for color images. Some researchers proposed to extract hand-crafted features from depth sequences by stacking shape and motion features over the whole video, and then convert them to texture images by encoding depth information. The generated texture images enable the use of existing models pre-trained on large scale image recognition or segmentation datasets with the fine-tuning operation to achieve satisfactory results. Wang *et al.* (2015b, 2016b) encoded the DMMs feature (Yang *et al.*, 2012a) into Pseudo-RGB images, which could convert the spatial and temporal movement information into textures and edges. Three independent ConvNets corresponding to three viewpoints were trained and the final recognition result was assigned by fusing the class score from three ConvNets. In (Wang *et al.*, 2016c), both the posture and motion feature in depth image sequences were extracted from dynamic depth images, dynamic depth normal images, and dynamic depth motion normal images. The spatial-temporal information was encoded into images using bidirectional rank pooling. These image-based representations enable existing CNN models trained on a large scale of image data for action recognition from depth sequences.

Rahmani & Mian (2016) proposed to learn a view-invariant human pose model from depth sequences. Each frame of real depth videos was input to the CNN model to learn a view-invariant and high-level feature space, and then new human poses captured from unknown views were transferred to this space to achieving a cross-view action recognition. These poses were clustered into different clusters and only the pose labels rather than action labels were used to learn the CNN model. They augmented the scale of multi-view training data by synthetically fitting 3D human models to real motion data and then produced several poses from different viewpoints. To overcome the limitations of skeleton data, such as a poor tolerance to self-occlusion and noise, Crabbe *et al.* (2015) proposed a skeleton-free body pose estimation from depth sequences by mapping the depth silhouette of the human body to the pose space through the CNN. They argued that this direct mapping without the intermediate skeleton step is more potential to be applied in general conditions.

2.5.3 Hybrid Feature-based Methods

Some researchers proposed to learn multi-modal features via separate networks for action recognition (Ijjina & Chalavadi, 2017; Miao *et al.*, 2017; Rahmani & Bennamoun, 2017; Wu *et al.*, 2016a; Zhang *et al.*, 2017a). Zhang *et al.* (2017a) proposed to use

3D convolutional neural networks (3DCNN) (Ji *et al.*, 2013; Tran *et al.*, 2015) and bidirectional convolutional long-short-term memory networks to learn spatio-temporal information from multi-modal data (RGB, depth, and flow produced from RGB). The final gesture recognition was achieved by combining learned multi-modal features in a linear SVM classifier. A Deep Dynamic Neural Networks (DDNN) (Wu *et al.*, 2016a) was designed for gesture recognition with multi-modal inputs. The DDNN includes a Gaussian-Bernoulli Deep Belief Network to explore dynamic features from skeleton sequences, and a 3DCNN to extract spatio-temporal features from RGB and depth images. This method proved that the fusion of multi-modal information could result in a better performance over uni-modal ones due to the complementary relation among different data channels.

Instead of the later fusion of results from each separate ConvNets, Scene Flow to Action Map (SFAM) was proposed to extract features from RGB and depth channels as one entity to ConvNets in (Wang *et al.*, 2017). Different variants of SFAM could encode effective spatial and temporal dynamics and enable the direct action recognition from two data modalities without later the score fusion. Alternatively, Shi & Kim (2017) investigated a privileged information-based RNN framework for action recognition from depth sequences. Skeleton joints provided during training was considered as a type of privileged information to achieve a better estimation of network parameters. Liu *et al.* (2016c) proposed to learn high-level features from raw depth images by designing a 3DCNN structure, while the low-level features such as the position and angle information between skeleton joints were calculated by the proposed JointVector. The classification results of SVM using both types of features were fused for the final action recognition. The model infers temporal information from the raw depth sequences by introducing the 3D filters.

2.5.4 Summary

This section reviewed different deep learning methods for human activity recognition using RGB-D data. According to the modality input to the neural networks, the methods were grouped into skeleton-based, depth-based, and hybrid feature-based. Most of the existing deep learning methods for action recognition rely on transforming the hand-crafted depth or skeleton features to texture images, and fine tune existing CNN models trained on larger scale image datasets for the classification of transformed images. The temporal information which is important to describe the evolution of an action is considered in different ways. A single image is captured by stacking the depth motion maps over a whole depth sequence in the majority of the depth-based methods. Some methods in the skeleton-based category restore the temporal information by converting skeleton features to a color image frame by frame. The different temporal duration problem is handled by resizing the final color image, which might introduce extra noise. LSTM is an effective solution to model the temporal information by introducing a gate mechanism to capture the temporal dependencies between frames.

2.6 Online Activity Recognition

Most of the human activity recognition methods rely on trimmed data provided by public datasets, however, the performance of these methods remain unknown when applied in online scenarios where the starting and ending time of the action are not given ahead. It is also necessary to detect activities with a low latency so the system can provide an instant response. For example, in human-robot interaction or collaboration scenarios, robots are expected to perform desirable activities by quickly interpreting human intentions. To localize the action, most of the early research uses a probability/energy-based threshold to detect the boundary or key poses of each action (Shan & Akella, 2014; Zhu *et al.*, 2016b). For example, Zhu *et al.* (2016b) identified transit motion features between two continuous poses in training phase, and the online classification was achieved by comparing likelihood probabilities in the Maximum Entropy Markov Model model.

There are some methods executing segmentation according to the clip-level or frame-level labelling approach (Devanne *et al.*, 2017a; Huang *et al.*, 2014; Kulkarni *et al.*, 2015; Wu *et al.*, 2015). Kulkarni *et al.* (2015) utilized DTW and dynamic frame warping for simultaneous action segmentation and classification. In their method, each video frame was assigned a label based on its comparison to the template representations, and the change of labels between consecutive frames indicated the starting or ending point of an event. Huang *et al.* (2014) developed sequential max-margin event detectors to spot an event from a continuous video with the presence of multiple event classes. The classifier was trained by using partial segments of events as the input. The most likely class was selected by penalizing other classes in the spot. Wu *et al.* (2015) clustered daily life clips to several action-words, with which an action-topics model was learned to reflect the co-occurrence and temporal relations. The action segmentation was realized according to the change of action topics between consecutive clips.

Sliding window is also a popular and compact technique for online action recognition (Gong *et al.*, 2014; Song *et al.*, 2012; Wu *et al.*, 2017), by which a video stream is usually divided into a set of overlapped segments and then classification is conducted in each segment. Similarly, in (Bloom *et al.*, 2013), an action label was allocated to each frame depending on the most confident prediction, and the boundary of actions in a video stream was determined by smoothing the calculated frame-level predictions via a sliding window. Such sliding window strategy has low computational efficiency and the difficulty of exploring a proper window size.

Instead of segmenting video streams using the sliding window strategy, Nowozin & Shotton (2012) proposed to detect action points which functioned as action peak frames to speed up the detection performance for online action recognition. Based on this, some approaches were developed for the action points detection in streaming videos (Bloom *et al.*, 2013, 2017; Fothergill *et al.*, 2012; Sharaf *et al.*, 2015). In (Sharaf *et al.*, 2015), the action peaks were identified according to action probabilities computed using a linear SVM classifier. A recursive feature elimination algorithm is proposed to select discriminative skeleton features whose covariance was then hierarchically encoded to represent human actions. Bloom *et al.* (2017) proposed to combine the clustered spatio-temporal manifolds and the temporal history of activities to detect the peaks of actions in a continuous stream. However, the detection result of a single time instance might not be representative enough for the complete action sequences and can cause false detections especially when the peak frames from different activities are quite similar.

More recently, some deep learning based methods address this problem by developing different architectures. Molchanov *et al.* (2016) proposed a recurrent 3DCNN to simultaneously perform classification and localization of hand gestures from continuous depth, color, and stereo-IR data sequences. They employed connectionist temporal classification (Graves *et al.*, 2006) to make gesture classification from the nucleus phase of the gesture without requiring particular segmentation. Shou *et al.* (2016) present to address action temporal localization via multi-stage CNNs, which includes identifying candidate segments that may contain actions, action recognition, and temporal boundary localization. Recently, many methods based on RNN or its variants (e.g., LSTM) have been proposed for online action recognition (Chai *et al.*, 2016; Li *et al.*, 2016; Song *et al.*, 2018), owing to their appealing capacity of modeling temporal dynamics of sequences. For example, Chai *et al.* (2016) proposed a spotting-recognition strategy to firstly segment continuous gestures into isolated gestures using the hand detector trained from Faster R-CNN, and then recognize the segmented gestures by fusing multi-modal features in a two streams RNN framework. Li *et al.* (2016) proposed a multi-task end-to-end Joint Classification Regression Recurrent Neural Network to simultaneously identify the action class and its temporal localization.

2.7 RGB-D Datasets

2.7.1 Human Action Datasets Using RGB-D Sensors

MSR-Action3D (Li *et al.*, 2010)

The MSR-Action3D dataset has 20 action types (Fig. 2.13 shows some samples), 10 subjects, and each subject performs each action for two or three times. The actions are: *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 swing, tennis serve, golf swing, and pickup throw*.

Florence3D-Action Dataset (Seidenari *et al.*, 2013)

This dataset has 9 actions and 10 subjects. Each action was performed two or three times: *wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch and bow.* Most of the actions, such as *answer phone* and *drink a bottle*, have a great similarity.

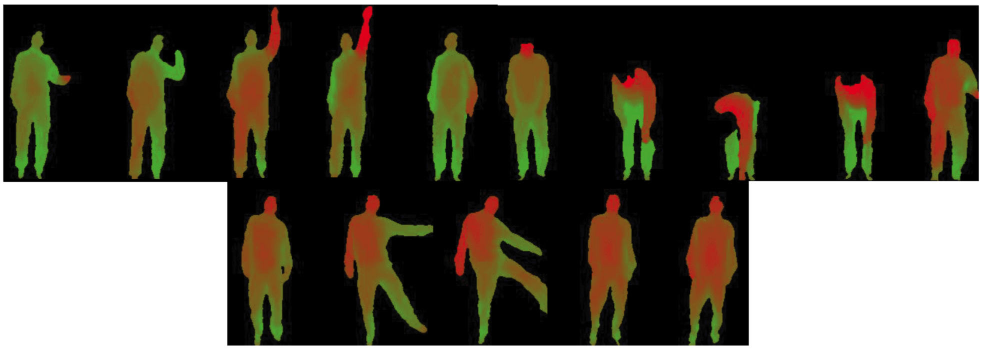


Fig. 2.13: Example frames of MSR-Action3D dataset (Li *et al.*, 2010).

MSRDailyActivity3D Dataset (Wang *et al.*, 2012b)

MSRDailyActivity3D Dataset aims to collect human's daily activities in the living room. It has 16 action types (Fig. 2.14 shows some samples), 10 subjects, and each subject performs each action for two times. The activities are: *drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down.* Some of the activities were performed in two different poses: *sitting on sofa* and *standing*. Compared to MSR-Action3D dataset, this dataset involves human-object interactions which makes it more challenging. This is due to the complex relation between human and objects is also needed to be considered to describe activities.

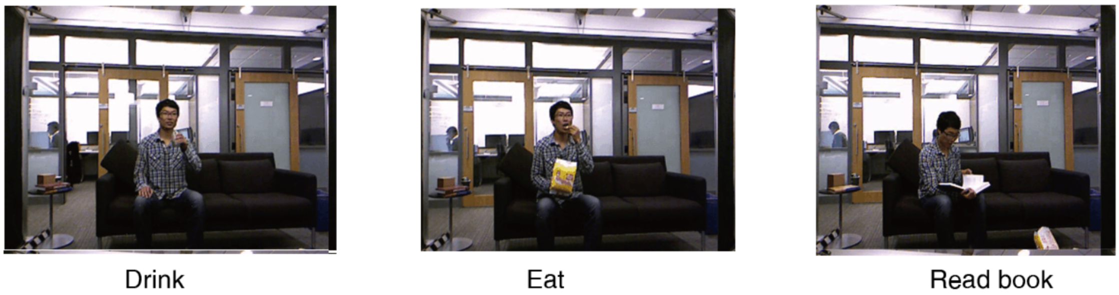
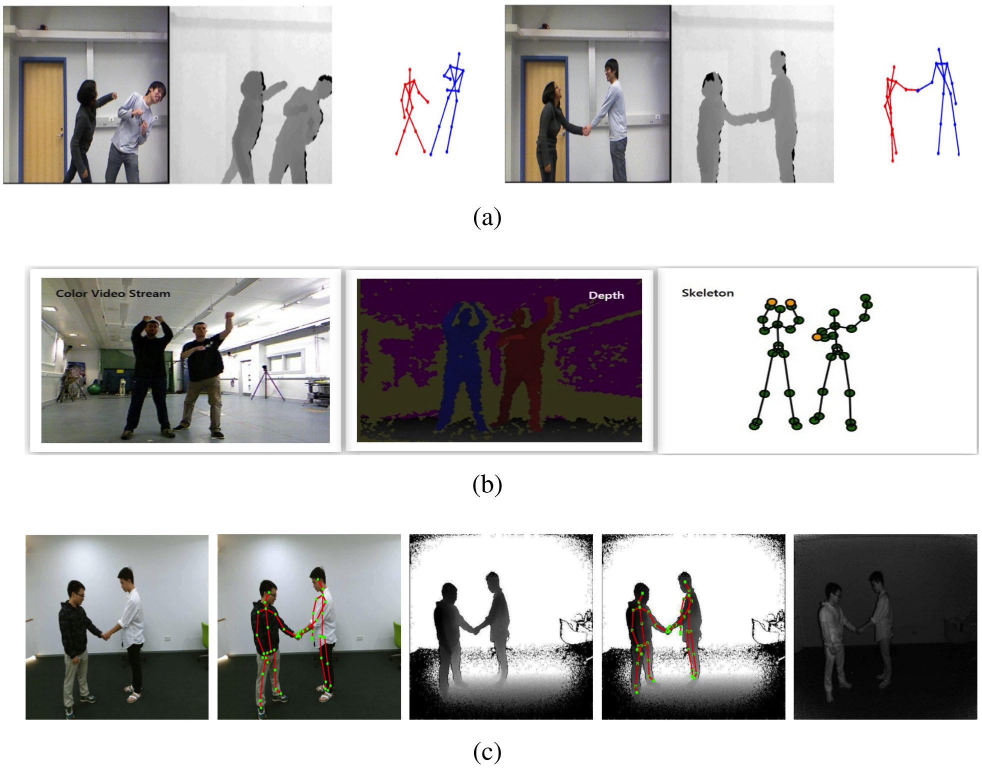


Fig. 2.14: Example frames of MSRDailyActivity3D dataset (Wang *et al.*, 2012b).

2.7.2 Human Interaction Datasets Using RGB-D Sensors

Although there are some early early human-human interaction datasets collected by RGB sensors, such as BEHAVE (Blunsden & Fisher, 2010), UT-Interaction (Ryoo & Aggarwal, 2010) and TV Human Interaction (Patron-Perez *et al.*, 2010), this subsection only reviews most of the publicly available RGB-D sensing based human interaction datasets. Fig. 2.15 shows some sample frames of each dataset.



(a)

(b)

(c)

Fig. 2.15: Example frames of (a) SBU Kinect Interaction dataset, (b) G3Di dataset, and (c) NTU RGB+D dataset.

SBU Kinect Interaction Dataset (Yun *et al.*, 2012)

This dataset is the first RGB-D dataset for human interaction. The dataset contains examples of 8 interaction classes: *approaching, departing, kicking, punching, pushing, hugging, shaking hands, exchanging something*, performed by 21 sets involving 7 participants. Compared to single person actions, these action categories are challenging because they are not non-periodic actions and have very similar body movements. This dataset is publicly used for the evaluation of recognition performance.

G3Di Dataset (Bloom *et al.*, 2014)

This dataset was collected for real time multi-player gaming. Interactions including collaborative and competitive mode in this dataset are virtual interactions during a computer game scenario. It has 6 sports: *boxing, volleyball, football, table tennis, sprint, hurdles*. Each sport involves various actions: *right punch, left punch, defend* for *boxing*, *serve, overhand hit, underhand hit, jump hit* for *volleyball*, *kick, block, save* for *football*, *serve, forehand hit, backhand hit* for *table tennis*, *run* for *sprint* and *run, jump* for *hurdles*. 12 people who were split into 6 pairs interacted through a computer interface.

NTU RGB+D Dataset (Shahroudy *et al.*, 2016a)

This dataset was collected in a varying indoor environment. Three cameras were set in different angles to capture each activity. To make the viewpoints manifold, the height and distances of the cameras to the subjects are alterable. The dataset has 60 activity classes including 40 daily actions, 9 health-related actions, and 11 human interactions. In comparison to the current datasets, the dataset is larger in classes, samples, and views due to variant camera settings.

2.7.3 Continuous Human Action Dataset

Multi-Modal Action Detection (MAD) Dataset (Huang *et al.*, 2014)

This is a sequential action database collected by Carnegie Mellon University in 2014. Fig. 2.16 shows some sample frames of the MAD database. The MAD database has 40 sequences performed by 20 subjects (2 sequences each subject). Each sequence contains 35 actions continuously performed by one subject. The time series between two actions are considered as the null class where the subject keeps standing in most cases. Three modalities: RGB videos, depth videos, and 3D coordinates of 20 skeleton joints were recorded using the Microsoft Kinect sensor.

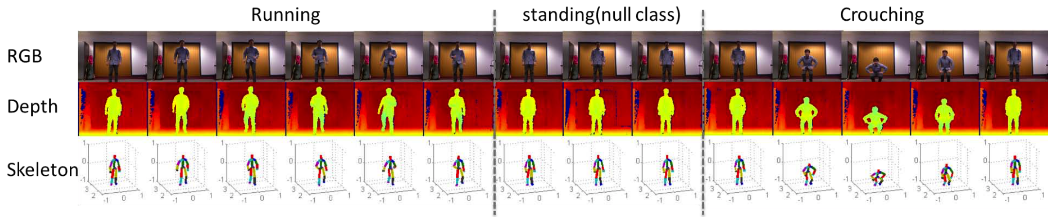


Fig. 2.16: Example frames of MAD dataset

PKU-MMD Dataset (Liu *et al.*, 2017b)

This is a large-scale dataset created for continuous human action understanding. Multi-modality including RGB, depth, infrared radiation and skeleton joints were recorded in this dataset. It has 51 action categories in total, which consist of 41 daily actions, such as *drinking, waving hand, combing hair, etc.*, and 10 interaction actions, such as *hugging, shaking hands, etc.* As the dataset is mainly aimed at action location and recognition from continuous sequences, it contains 1076 long video sequences, each of which contains about 20 action instances and lasts 3 to 4 minutes. 66 subjects of different ages were asked to perform actions.

2.7.4 Summary

This section reviewed different types of datasets collected using RGB-D sensors for human activity recognition. Most of the existing datasets may not be as challenging as realistic due to the involvement of constantly clustered or clean backgrounds. Moreover, most of the datasets provided only manually trimmed activity segments, which only contain one activity inside. This configuration does not mimic the practical scenario where activities are performed continuously. The big gap between the lab and practical scenarios makes it unclear of the online performance of the existing methods when applied in real-world. Therefore, the wild dataset without any constraints, where subjects acting actions naturally in the realistic environment, is imperative and promising for the future research.

2.8 Discussion

This section provides a discussion for both hand-crafted methods and deep learning methods in terms of feature types, adopted classifiers, and accuracies. Two commonly used human action datasets (MSR-Action3D and MSRDailyActivity3D) and one human interaction dataset (SBU Kinect Interaction) are selected for the comparison of different algorithms.

Table 2.1 categorizes techniques and compares their performance on two commonly used human action datasets. Each column of the table contains one type of methods, i.e., depth-based, skeleton-based or hybrid feature-based methods. Inside the column, the algorithms are further ranked according to the achieved accuracy. It can be seen that all the three categories of methods have achieved good recognition performance on the MSR Action 3D dataset due to its simplified experimental setting and action classes. Among them, 100% accuracy is obtained by Wang *et al.* (2016b) which converted the classic DMM to RGB images and utilized CNN for classification. However, their performance might decrease greatly in different viewpoint settings due to the dramatic variation of depth maps. Actually, based on the accuracy on this dataset, it is also easy to find that the skeleton-based methods are better suited for the classification of actions under different viewing angles than the depth-based methods and hybrid features-based methods. On the other hand, the hybrid features-based approaches outperform the skeleton-based or depth-based methods in the human-object interaction dataset of MSRDailyActivity3D, indicating that the skeleton alone is insufficient to distinguish actions which involve human-object interactions. The reason might be that the context information of objects also plays an important role in the defined actions.

Table 2.1 also divides the methods into hand-crafted methods and deep learning methods. The table shows that the top recognition accuracy of MSR Action3D dataset is achieved by deep learning based methods, which demonstrates their effectiveness in human action recognition. Compared to the former dataset, fewer deep learning based methods are evaluated on the MSRDaliyAcitivity3D dataset and hand-crafted methods achieve better performance at this stage. Regarding the classifier, Most of the hand-crafted methods adopt the SVM, while deep learning methods normally use CNN, LSTM or their combination for recognition.

Table 2.1: Recognition performance of the state-of-the-art methods on *MSR Action3D* and *MSRDailyActivity3D*.

Notation: Ref.: Reference; H-C: Hand-Crafted methods; DL: Deep Learning methods; PDF: probability density function; RF: Random Forest; Acc.: Recognition accuracy (%).

	MSR Action3D-following evaluation protocol Li *et al.* (2010)								
	Depth-based			**Skeleton-based**			**Hybrid feature-based**		
	Ref.	Classifier	Acc.	Ref.	Classifier	Acc.	Ref.	Classifier	Acc.
H-C	Xia & Aggarwal (2013)	SVM	89.30	Vemulapalli *et al.* (2014)	SVM	92.46	Wang *et al.* (2014)	SVM	88.20
	Devanne *et al.* (2015)	kNN	92.10	Theodorakopoulos *et al.* (2014)	kNN	93.61	Ji *et al.* (2018)	SVM	90.80
	Yang & Tian (2014b)	SVM	93.90	Koniusz *et al.* (2016)	SVM	93.96	Jalal *et al.* (2017)	HMM	93.30
	Liu & Liu (2016)	SVM	**94.28**	Wang *et al.* (2016a)	Matching	94.40	Kong *et al.* (2016)	SVM	93.99
				Liu *et al.* (2016a)	SVM	94.40	Zhu *et al.* (2013)	RF	94.30
				Guo *et al.* (2018)	SVM	**95.24**	Ohn-Bar & Trivedi (2013)	SVM	**94.84**
DL	Wang *et al.* (2015c)	CNN	94.58	Veeriah *et al.* (2015)	RNN	92.03	Liu *et al.* (2016c)	CNN	84.07
	Wang *et al.* (2016b)	CNN	**100.0**	Du *et al.* (2015b)	RNN	94.49	Kamel *et al.* (2018)	CNN	94.51
				Núñez *et al.* (2018)	CNN+LSTM	**96.00**	Shi & Kim (2017)	RNN	**94.90**
	MSRDailyActivity3D-following evaluation protocol Wang *et al.* (2012b)								
	Depth-based			**Skeleton-based**			**Hybrid feature-based**		
	Ref.	Classifier	Acc.	Ref.	Classifier	Acc.	Ref.	Classifier	Acc.
H-C	Oreifej & Liu (2013)	SVM	80.00	Zanfir *et al.* (2013)	kNN	73.80	Kong *et al.* (2016)	SVM	73.21
	Yang & Tian (2014b)	SVM	86.25	Qiao *et al.* (2017)	SVM	75.00	Ji *et al.* (2018)	SVM	81.30
	Jia & Fu (2016)	SVM	80.63	Cai *et al.* (2016)	MIL	78.52	Zhang & Parker (2016)	SVM	86.00
	Chen *et al.* (2017)	ELM	**89.00**	Liu *et al.* (2017d)	SVM	**91.00**	Wu *et al.* (2016b)	SVM	87.20
							Shahroudy *et al.* (2016b)	SVM	91.25
							Althloothi *et al.* (2014)	SVM	93.10
							Jalal *et al.* (2017)	HMM	**94.10**
DL	Wang *et al.* (2015c)	CNN	78.12	Núñez *et al.* (2018)	CNN+LSTM	**63.10**			
	Wang *et al.* (2016b)	CNN	85.00						
	Luo *et al.* (2017)	CNN+LSTM	**86.90**						

Table 2.2: Comparison of the state-of-the-art methods on *SBU Kinect Interaction*.

Notation: Separately: consider each person's action as an individual sample and averaging the classification scores for the final prediction; Concatenation: simply stack each person's feature together or further include the interrelationship between the two persons.

SBU Kinect Interaction(5-fold cross-validation Yun *et al.* (2012))				
Reference	Feature	Accuracy (%)	Type	Interaction solution
Yun *et al.* (2012)	Skeleton	80.30	MIL	Distance
Ji *et al.* (2014)	Skeleton	86.90	SVM	Body part
Ji *et al.* (2015)	Skeleton	89.40	SVM	Body part
Wu *et al.* (2017)	Skeleton	91.00	-	Distance
Liu *et al.* (2017a)	Skeleton	**91.12**	SVM	Body part
Reference	Feature	Accuracy (%)	Type	Interaction solution
Zhu *et al.* (2016c)	Skeleton	90.4	LSTM	Seperately
Song *et al.* (2017)	Skeleton	91.5	LSTM	Concatenation
Liu *et al.* (2016b)	Skeleton	93.3	LSTM	Concatenation
Ke *et al.* (2017)	Skeleton	93.6	CNN	Seperately
Wang & Wang (2017)	Skeleton	94.8	RNN	Seperately
Liu *et al.* (2018)	Skeleton	94.90	LSTM	Concatenation
Zhang *et al.* (2017b)	Skeleton	**97.2**	RNN	Concatenation

Table 2.2 reports a comparison of the state-of-the-art methods on the commonly used human interaction dataset: SBU Kinect Interaction, in terms of accuracies, feature types, classifiers, and solutions to the interaction challenge. The interaction challenge lies in the adapting of single human's action features into a representation that is suitable for the human interaction scenario. As shown in the table, the existing solutions can be grouped into four categories (joints distance, interactive body part, separately, concatenation) whose definition is shown in the caption of the table. The top performance of deep learning based methods (97.2%, Zhang *et al.* (2017b)) outperforms the top hand-crafted based method (91.12%, Liu *et al.* (2017a)) to a large extent. It can also be observed that all of the human interaction approaches are based on the skeleton data. In hand-crafted methods, the inter-relationship between two persons is modeled by using interactive body parts or joints distance information. While in the deep learning based methods, this challenge is handled by a concatenation operation or the simple separation strategy which recognizes each person's action and averages the classification scores for the final prediction.

It can be observed from the tables that deep learning based methods have achieved superior recognition performance over hand-crafted based methods in most of the existing human activity datasets. However, it is also well-known that most of the deep learning based approaches require large training samples to reduce the affect of over-fitting and achieve better performance. Apart from the recognition accuracy, it is also essential for the algorithms to be computationally efficient for many real-world motion recognition applications. Most of the existing hand-crafted methods achieved real-time performance via a careful design of the features and the use of low computational cost classifiers such as SVM. Due to the complex structure of neural networks, existing deep learning based methods heavily rely on advanced parallel computing devices such as GPU and TPU to reach real time performance.

Chapter 3

Spatio-temporal Representation for Human Action Recognition

3.1 Introduction

The emergence of cost-efficient RGB-D sensors eases the difficulties of the high sensitivity to illumination conditions and texture variability of the RGB data (Niebles & Fei-Fei, 2007; Niebles *et al.*, 2008) and reveals a promising direction for human activity recognition by providing depth data. The depth information also enables 3D human skeleton joints to be easily estimated (Shotton *et al.*, 2013b). Fig. 3.1 shows the configuration of 20 body joints. A large number of research has been done for human action recognition using skeleton data. Various characteristics of skeleton joints, such as locations, angles, and geometric relationships, were utilized to model different human actions (Gaglio *et al.*, 2015; Gowayyed *et al.*, 2013; Lillo *et al.*, 2017; Pazhoumand-Dar *et al.*, 2015; Qiao *et al.*, 2017). However, accurate recognition still remains a challenge because of various object appearances, poses, and video sequences.

To achieve effective human action recognition, this chapter proposes a spatio-temporal feature descriptor to describe various human actions. Specifically, in spatial domain, a kernel enhanced BSW for each joint is constructed to represent the moving trend of skeleton joints using an effective histogram projection method. The directions in BSW are grouped into semantic moving words, whose distribution over an activity sequence explicitly interprets the moving trend of skeleton joints. Furthermore, the

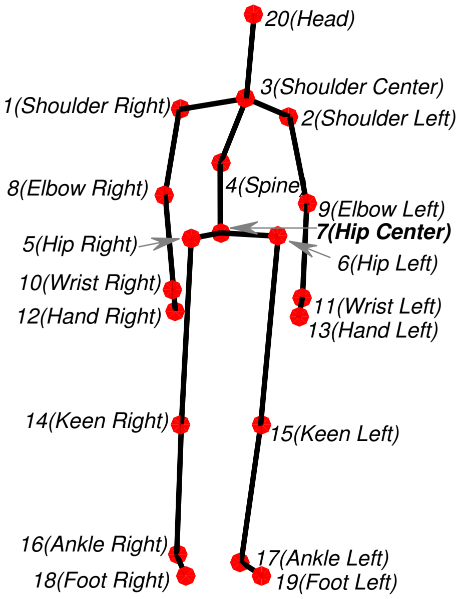


Fig. 3.1: Configuration of 20 body joints.

kernel-based dynamic weighting strategy is developed to augment the informative features. This feature could describe the specific tendency of skeleton joints in 3D space. In temporal domain, the geometry information of joints in each frame is modeled by the relative motion with the initial status.

The remainder of this chapter is organized as follows: Section 3.2 introduces the proposed human action representation. Section 3.3 reports experimental results as well as the comparison with the state-of-the-art methods. Section 3.4 summarizes the work of this chapter.

3.2 Feature Extraction

3.2.1 Data Pre-processing

The original coordinate of joints obtained by Kinect sensors is translated to the proposed person-centric coordinate system (as shown in Fig. 3.2). This coordinate transformation makes features invariant to various locations and orientations by extracting

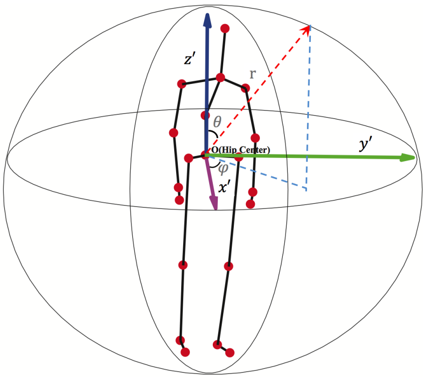


Fig. 3.2: Person-centric coordinate system.

them in the relative position rather than the absolute position. It is defined as follows: the z' axis can be calculated using the vector from the hip center to the spine joint, and its unit vector is denoted as (a_7, a_8, a_9); the x' axis is the normal vector of a plane constructed by the spine point, left hip point and right hip point, and its unit normal vector is represented by (a_1, a_2, a_3); finally, the y' axis can be determined by the dot product of above two unit vectors, and the value of its unit y' is (a_3, a_4, a_5). Consequently, the transformation of coordinates is calculated using the following equation:

$$P = R * P' + T \tag{3.1}$$

where P and P' denote the original coordinate and the transformed coordinate, respectively, and T is the coordinate of the hip center $[x_h, y_h, z_h]^{-1}$. R is the rotation matrix:

$$R = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & a_9 \end{bmatrix}^{-1}$$

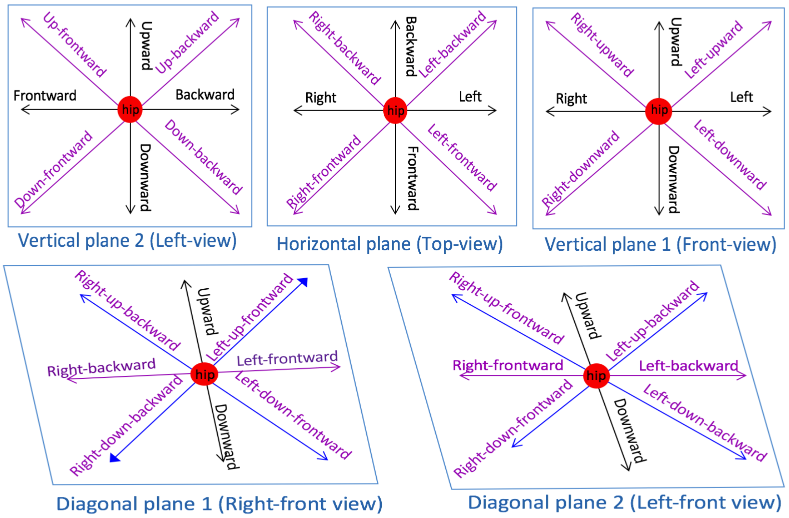
3.2.2 3D Moving Trend Feature

Semantic Moving Words

The moving directions of body joints in 3D space are various while actors performing different actions, therefore, the directions in 3D space are divided into several semantic moving words $\mathbf{V_w}$, and a distribution of movements over these semantic moving words is captured to interpret the moving trend of an activity sequence. The criterion in selecting the size of the semantic moving words is based on whether the constructed moving words can equally divide the 3D space and it is representative for the moving direction. The effect of different sizes of the semantic moving words (ranging from 6, 14, 26, 42, 62, 86 to 114) on the recognition result will be studied in the experiment section. Let $\mathbf{V_w} = [\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_m}]$ be the matrix of m words in 3D space (here, $m = 26$ is considered for example), which are given by:

$$\begin{array}{lll}
\mathbf{v_1} = (0,0,1)^T, & \mathbf{v_2} = (0,0,-1)^T, & \mathbf{v_3} = (0,1,0)^T, \\
\mathbf{v_4} = (0,-1,0)^T, & \mathbf{v_5} = (1,0,0)^T, & \mathbf{v_6} = (-1,0,0)^T, \\
\mathbf{v_7} = (1,1,1)^T, & \mathbf{v_8} = (-1,-1,-1)^T, & \mathbf{v_9} = (1,1,-1)^T, \\
\mathbf{v_{10}} = (-1,-1,1)^T, & \mathbf{v_{11}} = (1,-1,1)^T, & \mathbf{v_{12}} = (-1,1,-1)^T, \\
\mathbf{v_{13}} = (1,-1,-1)^T, & \mathbf{v_{14}} = (-1,1,1)^T, & \mathbf{v_{15}} = (1,1,0)^T, \\
\mathbf{v_{16}} = (-1,-1,0)^T, & \mathbf{v_{17}} = (1,-1,0)^T, & \mathbf{v_{18}} = (-1,1,0)^T, \\
\mathbf{v_{19}} = (-1,0,-1)^T, & \mathbf{v_{20}} = (1,0,1)^T, & \mathbf{v_{21}} = (1,0,-1)^T, \\
\mathbf{v_{22}} = (-1,0,1)^T, & \mathbf{v_{23}} = (0,1,1)^T, & \mathbf{v_{24}} = (0,-1,-1)^T, \\
\mathbf{v_{25}} = (0,1,-1)^T, & \mathbf{v_{26}} = (0,-1,1)^T &
\end{array} \tag{3.2}$$

Fig. 3.3 shows the defiend 26 semantic words in 3D space.



Vertical plane 2 (Left-view)
Horizontal plane (Top-view)
Vertical plane 1 (Front-view)
Diagonal plane 1 (Right-front view)
Diagonal plane 2 (Left-front view)

Fig. 3.3: Samples of semantic moving words (taking $m = 26$ for example).

To augment the discriminative information of skeleton joints, a kernel enhanced

BSW is proposed, where features in both spatial and temporal domain are jointly weighted based on their contribution to an activity. As shown in Fig. 3.4, the moving trend of joints is captured from the front, side, and top view. The moving characteristic

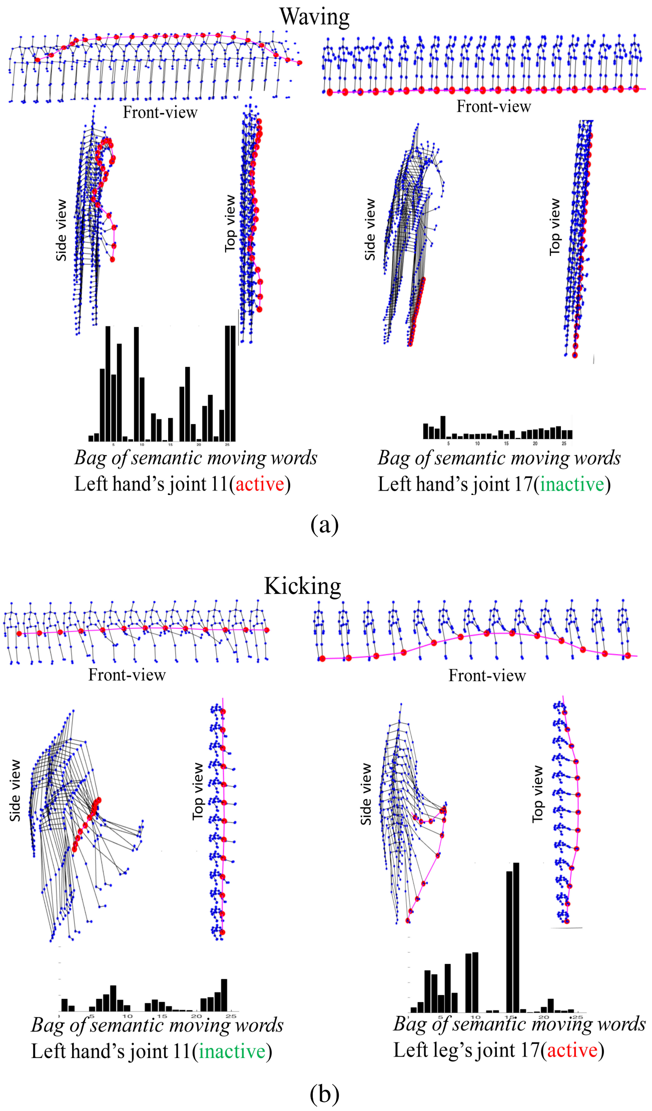


Fig. 3.4: The comparison of moving trends of skeleton joints in actions *waving* and *kicking*.

for the joints in the left hand (11) and leg (17) captured from the action *waving* and *kicking* are shown in (a) and (b) respectively. Joint 11 with apparent moving property is regarded as the active one in the action *waving*, where joint 17 with few move is inactive. While joint 11 is inactive in the action *kicking* where joint 17 is opposite. To conclude, the moving trend of the active joint is more apparent than that of the inactive joint, while the moving trend of the same joint in different classes is also diverse. Thus, it is reasonable to use the moving trend of skeleton joints to discriminate different action categories.

For the $i-th$ joint, given a point set:

$$P^i = \{p_1^i, ..., p_t^i, ..., p_F^i\} \tag{3.3}$$

where F denotes the length of action sequence, and t denotes the time. Since p_t^i includes three coordinates x, y, z. The 3D direction vector $\mathbf{v_t^i}$ of the $i\ th$ joint is obtained via p_t^i and p_{t-1}^i:

$$\mathbf{v_t^i} = \{x_{p_t^i} - x_{p_{t-1}^i}, y_{p_t^i} - y_{p_{t-1}^i}, z_{p_t^i} - z_{p_{t-1}^i}\} \tag{3.4}$$

and then calculate the $cos\langle \mathbf{v_t^i}, \mathbf{v_j} \rangle$ of angle $\theta^i(t)$ between $\mathbf{v_t^i}$ and m vectors:

$$cos\theta_j^i(t) = \frac{\mathbf{v_j} \cdot \mathbf{v_t^i}}{\|\mathbf{v_t^i}\| \|\mathbf{v_j}\|}, j \in [1, m] \tag{3.5}$$

where $\mathbf{v_j} \in \mathbf{V}$. Since the greater the $cos\theta_j^i(t)$ value, the more similar the direction Singhal *et al.* (2001), the cosine similarity $cos\theta_j^i(t)$ is calculated to describe the similarity between $\mathbf{v_t^i}$ and $\mathbf{v_j}$.

Bag of Semantic Moving Words

(Van Gemert *et al.*, 2010) showed that the image classification performance of the distribution of visual words using the soft-assignment strategy is superior than that of the hard-assignment strategy. Instead of distributing the probability over all words, a Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}\delta} exp(-\frac{(x-\mu)^2}{2\delta^2})$ is applied for the soft voting of the moving trend histogram. This soft voting has the capacity of weighting salient motion during building the final histogram. Since the *cosine* similarity is suited to measure the distance between two direction vectors, a Gaussian kernel function is improved by considering it as the variable. The weight during the voting is then computed as follows:

$$K(cos(\mathbf{v}_t, \mathbf{v}_w^i)) = \frac{1}{\sqrt{2\pi}\delta} exp(-\frac{(cos(\mathbf{v}_t, \mathbf{v}_w^i)) - \mu)^2}{2\delta^2}) \tag{3.6}$$

where μ and δ are the mean and standard deviation of *cosine* values, respectively. Here, the mean of the Gaussian function is 1 due to the trait of *cosine* values.

A soft-assignment strategy where a frame direction is distributed to multiple most relevant word candidates is achieved using a $1 \times f$ vector S. The soft voting degree is controlled by using a parameter k to determine the elements in S. For example, if $k = 3$, the frame direction is encoded to the 3 words with the top 3 similarities. Thus, the elements of S satisfy:

$$S_t = \begin{cases} 1 & \textit{if the value } cos(\mathbf{v}_t, \mathbf{v}_w^i) \textit{ belongs to} \\ & \quad \textit{k biggest similarities} \\ 0 & otherwise \end{cases} \tag{3.7}$$

To weight the frames which make bigger contributions to the whole sequence, the frame displacement $Dis(t) = \|\mathbf{v_t}\|$ is added during a quantization process. Therefore, the frame weight function can be achieved as follows:

$$w(t) = Dis(t) * K(cos(\mathbf{v}_t, \mathbf{v}_w^i)) \tag{3.8}$$

The final representation of each word is built by accumulating the movement through the action sequence:

$$BSW(\mathbf{v}_w^i) = \sum_{t=1}^{f} S_t * w(t) \tag{3.9}$$

3.2.3 Geometry Property

To remove the coordinate difference caused by various distances between actors and the depth sensor, the world coordinate from the depth sensor is translated to the center of actors in each frame. Although the world coordinate of each frame may differ under current strategy, the advantage is obvious as hip-center point is relatively stable in majority of actions. Apart from the feature of hip-center point relative movement, it should be noted that different actors might have different initial poses for the same action. In order to eliminate the influence of different initial poses for the rest 19 joints, the displacement between the relative joints in current frame and the joints in the initial frame is applied to reflect the geometry property in current frame.

Furthermore, the action recognition performance is affected by the various body sizes of the actors. This is caused by internal difference of human or various distances between actors and the depth sensor. To solve this problem, a feature normalization method is performed on the extracted geometry property feature. The movement of a point can be regarded as the composition of movements of x, y, z axes. In each single frame, the relative motion of each joint to its initial status in three axes is recorded. Each frame represents a body pose that can be described by the locations of 20 joints.

$$I_t = \{p_t^1, p_t^2, ..., p_t^N\} \tag{3.10}$$

where N is the number of joints, p_t^i is the position of the $i-th$ joint at time t and it contains 3D coordinates x_t^i, y_t^i and z_t^i. The difference along three axes of each joint can be computed between the initial status and the current status.

Firstly, the world coordinate system is translated to the hip-center joint for each frame using Eq. 3.1. The transformed coordinates of skeleton joints are denoted as p_t^{ri}. So the transformed coordinates of the frame is $I_{rt} = \{p_t^{r1}, p_t^{r2}, ..., p_t^{rN}\}$ and the geometry property of each joint at frame t is denoted as:

$$\begin{cases} \triangle x_t^i = x_t^{ri} - x_1^{ri}, \\ \triangle y_t^i = y_t^{ri} - y_1^{ri}, \\ \triangle z_t^i = z_t^{ri} - z_1^{ri}, \end{cases} \tag{3.11}$$

where $(x_1^{ri}, y_1^{ri}, z_1^{ri})$ and $(x_t^{ri}, y_t^{ri}, z_t^{ri})$ are the three transformed coordinates of the initial status and current status, respectively. The relative displacement of the $i-th$ joint at frame t is $\triangle d_t^i : (\triangle x_t^i, \triangle y_t^i, \triangle z_t^i)$, and the geometric property of current frame is:

$$g(t) = \{\triangle d_t^1, ..., \triangle d_t^N\} \tag{3.12}$$

$G(k) = \{g(1), ..., g(F)\}$ denotes the feature of action k. So the dimension of the defined geometric property feature for one frame is 20×3. Although (Yang & Tian, 2012) also uses the difference of the joints between current frame and the initial frame, the proposed geometric property feature is totally different. In (Yang & Tian, 2012) different combination of the joints is used and the final dimensions for each frame is 400×3, which is 20 times larger than the proposed feature dimensions.

The length of action sequences may differ in each action instance, and this will lead to unequal length of the geometry property feature. Therefore, the extracted feature is rescaled using the cubic spline interpolation (Vemulapalli *et al.*, 2014) before integrating them into the feature descriptor.

Finally, to acquire the scale-invariant $G(k)$ feature for the different body sizes, the following normalization method is used:

$$G_n(k) = \frac{G(k)}{\parallel G(k) \parallel} \tag{3.13}$$

3.2.4 GBSW Representation

The general framework of the proposed GBSW representation is shown in Fig. 3.5. The upper part of Fig. 3.5 is the proposed BSW where a histogram of 26 bins corresponding to 3D moving directions is adopted to store the moving trend of each joint through the whole action video. The lower part of Fig. 3.5 is the G feature which is acquired from the N frames of the action sequence. In the G feature, the world coordinate is firstly translated into hip-center using Eq.(3.3) and the relative displacement of each joint is computed by using Eq.(3.11). To address unequal length of the G feature caused by different length of action sequences, the relative displacement property of each action instance is interpolated to the unified dimension, $M \times 20 \times 3$.

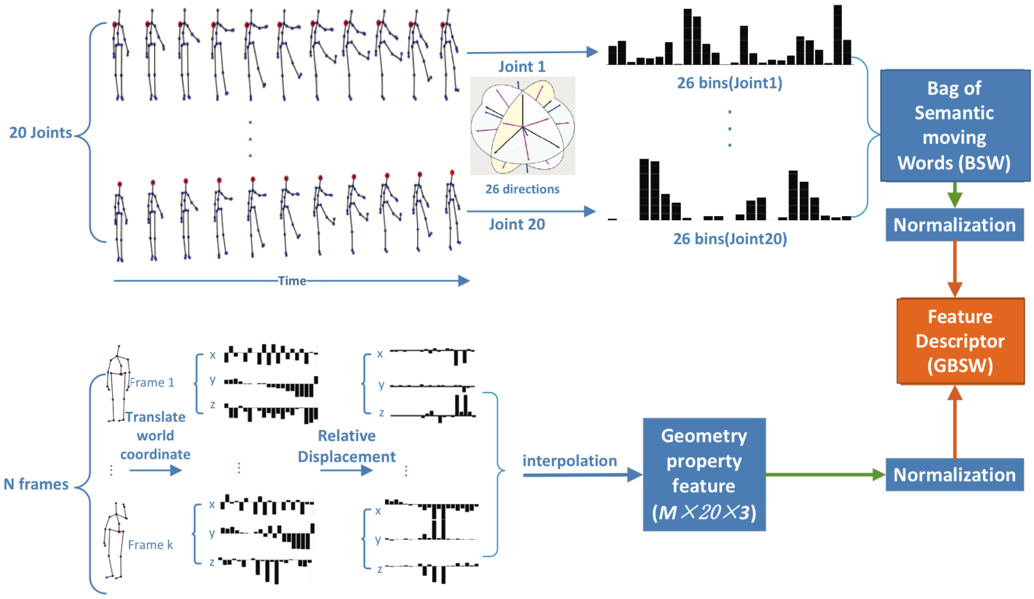


Fig. 3.5: An overview of the proposed GBSW feature descriptor.

The final feature descriptor GBSW is a concatenation of the G feature and BSW feature. The G feature indicates the temporal movement of each joint, while the BSW feature reflects spatial motion direction of each joint in an action sequence.

3.3 Experimental Evaluation

3.3.1 Introduction of Datasets

This subsection gives detailed information about the adopted RGB-D datasets (MSR-Action3D Dataset and Florence3D-Action Dataset) which are commonly used to compare the performance of human action recognition algorithms. The experiment on the two datasets aims to demonstrate the comparative results of the proposed GBSW on human action recognition.

MSR-Action3D dataset (Li *et al.*, 2010) has 20 human action categories. The data is divided into three action sets *AS1*, *AS2* and *AS3*, as shown in Table 3.1. Actions with similar movement are grouped in the *AS1* and *AS2* sets, while complex actions are grouped in *AS3* set. Each set has eight actions with some overlaps between action sets. In each action set, there are three tests with different settings of training and testing samples: **Test One**: 1/3 of the samples for training and the rest for testing; **Test Two**: 2/3 of the samples for training and the rest for testing; **Cross Subject Test**: samples from half of subjects for training and the rest for testing. To carry out a fair comparison,

Table 3.1: Three action sets of *MSR-Action3D* dataset.

AS1	*AS2*	*AS3*
Horizontal Wave	High Wave	High Throw
Hammer	Hand Catch	Forward Kick
Forward Punch	Draw X	Side Kick
High Throw	Draw Tick	Jogging
Hand Clap	Draw Circle	Tennis Swing
Bend	Hands Wave	Tennis Serve
Tennis Serve	Forward Kick	Golf Swing
Pickup & Throw	Side Boxing	Pickup & Throw

experiment settings following the protocols from (Li *et al.*, 2010) (samples of subject 1, 3, 5, 6, 9 as training) and (Zanfir *et al.*, 2013) (samples of subjec 1, 2, 3, 4, 5 as training) in *Cross Subject Test* are conducted.

For Florence3D-Action dataset, the *Cross Subject Test* setting from (Vemulapalli *et al.*, 2014) is used for the performance comparison.

3.3.2 Evaluation of the GBSW Representation

Having computed above feature descriptors, a linear SVM (Chang & Lin, 2011) algorithm is then applied for action classification. To show the superior performance of the combination of different features, the comparison between its recognition result and that of the single features on two datasets is listed in Table 3.2. It can be seen that the GBSW feature which combines geometric property and bag of semantic moving words improves the performance in each experimental setting, which proves that the specific information from different types of features can complement each other. For example, the geometric feature seems to be complementary in term of spatial information to the motion feature in BSW, which enables the hybrid representation to be more discriminative among different activity categories.

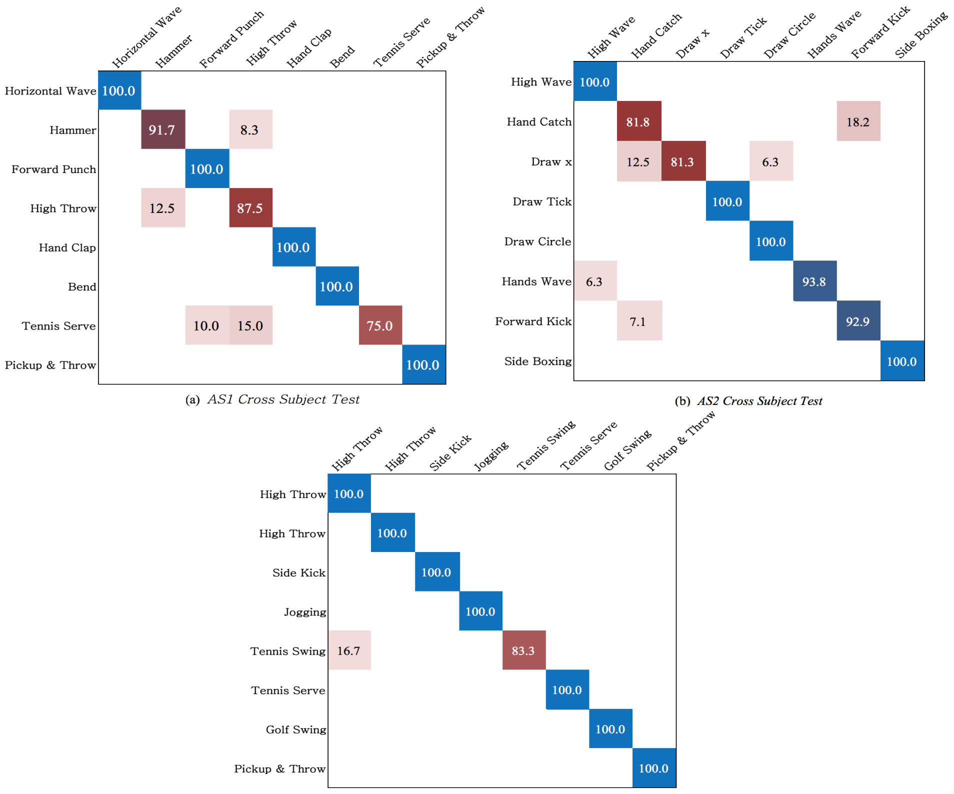
Table 3.2: Recognition accuracy (%) of different features on *MSR-Action3D* and *Florence3D-Action* dataset

Feature type	*MSR-Action3D*			*Florence3D-Action*
	AS1	*AS2*	*AS3*	*Cross Subject Test*
G	50	79.5	92.4	85.9
BSW	92.4	85.7	93.3	88.0
GBSW	**93.4**	**94.9**	**98.4**	**93.6**

In addition, to evaluate the effect of the number of semantic moving words (n_s), the recognition performance of the proposed method with $n_s = 6, 14, 26, 42, 62, 86, 114$ is shown in Table 3.3. The recognition accuracy increases till $n_s = 26$, while it decreases when n_s is over 26. This is because the rising number of semantic moving words augments the ambiguous moving trend between actions, which influences the discriminating capacity of the feature. Based on this finding, $n_s = 26$ is selected to get the following performance.

Table 3.3: The average recognition accuracy (%) of the proposed feature representation versus the size of semantic moving words.

Size of moving words	*MSR-Action3D*	*Florence3D-Action*
6	90.03	89.30
14	92.26	91.10
26	**94.40**	**91.30**
42	93.45	89.53
62	93.45	88.95
86	94.40	89.90
114	92.23	89.00



(a) *AS1 Cross Subject Test*

(b) *AS2 Cross Subject Test*

(c) *AS3 Cross Subject Test*

Fig. 3.6: Confusion Matrixes of the proposed GBSW feature descriptor: *AS1*,*AS2* and *AS3* on MSR-Action3D dataset.

For *MSR-Action3D* dataset, Fig. 3.6 shows the confusion matrices of *AS1*, *AS2* and *AS3* on *Cross Subject Test* achieved using the proposed GBSW. It can be seen that most actions can be 100% recognized by the proposed descriptor, especially for *AS3* where all actions except *Tennis Swing* are correctly recognized. Because actions in *AS1* and *AS2* have big inter-class similarity, some actions are similar with others, such as *Hammer* and *High Throw*, *Tennis Serve* and *Forward Punch*. As a result, the recognition accuracies of these actions are lower than those actions with smaller intra-class variations.

Table 3.4: Recognition Accuracy (%) of *Test One* and *Test Two* on *MSR-Action3D*.

Test One				
Methods	*AS1*	*AS2*	*AS3*	**Average**
Bag of 3D Points (Li *et al.*, 2010)	89.5	89.0	96.3	91.6
DMM-HOG (Yang *et al.*, 2012b)	97.3	92.2	98.0	95.8
STOP (Vieira *et al.*, 2014)	**98.2**	94.8	97.4	96.8
HOJ3D (Xia *et al.*, 2012)	98.5	96.7	93.5	96.2
EigenJoints (Yang & Tian, 2012)	94.7	95.4	97.3	95.8
(Jalal *et al.*, 2017)	96.9	**98.3**	**98.7**	97.9
3GMTG (Liu *et al.*, 2016a)	94.7	95.0	96.8	95.5
GBSW	97.9	98.2	98.5	**98.2**
Test Two				
Methods	*AS1*	*AS2*	*AS3*	**Average**
Bag of 3D Points (Li *et al.*, 2010)	93.4	92.9	96.3	94.2
DMM-HOG(Yang *et al.*, 2012b)	98.7	94.7	98.7	97.4
STOP (Vieira *et al.*, 2014)	**99.1**	97.0	98.7	98.3
HOJ3D (Xia *et al.*, 2012)	98.6	97.9	94.9	97.2
EigenJoints (Yang & Tian, 2012)	97.3	**98.7**	97.3	97.8
(Jalal *et al.*, 2017)	97.1	98.6	98.9	98.2
3GMTG (Liu *et al.*, 2016a)	98.5	97.8	**99.1**	98.5
GBSW	98.2	**98.7**	**99.1**	**98.7**

3.3.3 Comparison with State-of-the-art Methods

The following subsections present a comparison of the proposed method with the state-of-the-art methods in terms of recognition accuracy.

Table 3.4 reports the results of *Test One* and *Test two* on the *MSR-Action3D* dataset. It can be seen that the GBSW representation obtains highest average recognition rates in both cases. Specifically, the proposed method achieved better performances on three action sets than the skeleton-based methods, such as HOJ3D (Xia *et al.*, 2012) and EigenJoints (Yang & Tian, 2012). It should be noted that the highest accuracy of *AS1* achieved by STOP (Vieira *et al.*, 2014), which jointly uses the skeleton and depth information, indicates that the recognition performance of similar actions might be improved by an effective fusion of the depth information.

To evaluate the adaptability of the proposed method across subjects, the experiment results on *Cross Subject Test* are reported in Table 3.5. To execute a fair comparison, the considered methods are sorted into groups according to two protocols from (Li *et al.*, 2010) and (Zanfir *et al.*, 2013), respectively. The compared methods on *MSR-Action3D* dataset are further categorized into silhouette-based (Li *et al.*, 2010; Oreifej & Liu, 2013; Yang & Tian, 2014b; Yang *et al.*, 2012b), local interest points-based (Vieira *et al.*, 2014; Wang *et al.*, 2012a; Xia & Aggarwal, 2013) and skeleton-based (Devanne *et al.*, 2015; Vemulapalli *et al.*, 2014; Wang *et al.*, 2013, 2012b; Xia *et al.*, 2012; Yang & Tian, 2012; Yang *et al.*, 2017a,b; Zanfir *et al.*, 2013) in Table 3.5. The GBSW method obtained recognition rates over 90% on *AS1*, *AS2* and *AS3* with the procotol from (Li *et al.*, 2010) and the rates were over 95% on *AS2* and *AS3* with the procotol from (Zanfir *et al.*, 2013). The proposed method outperformed silhouette-based methods, for example, the recognition accuracy of GBSW was approximately 20% higher than that of (Li *et al.*, 2010). Moreover, the proposed descriptor improved the average recognition rate by 6.3% compared to DSTIP (Xia & Aggarwal, 2013) which is the best result of the listed local interest points-based methods.

In addition, Table 3.6 records the *Cross Subject Test* performance of different methods on *Florence3D-Action* dataset. Some actions in this dataset are quite confused with each other, for example, the body movement in *answer phone* and *drink a bottle* is similar. The table shows that the proposed feature descriptor performed 93.6% recognition

Table 3.5: Average accuracy (%) of *Cross Subject Test* on the *MSR-Action3D*.(1 silhouette-based, 2 local interest points-based, 3 skeleton-based)

Protocol from (Li *et al.*, 2010) (samples of subject 1, 3, 5, 6, 9 as training)					
	Method	*AS1*	*AS2*	*AS3*	**Average(%)**
1	Bag of 3D Points(Li *et al.*, 2010)	72.9	71.9	79.2	74.7
	DMM-HOG(Yang *et al.*, 2012b)	96.2	84.1	94.6	91.6
	SNV(Yang & Tian, 2014b)	-	-	-	93.1
2	STOP (Vieira *et al.*, 2014)	91.7	72.2	98.6	87.5
	ROP (Wang *et al.*, 2012a)	-	-	-	86.5
	DSTIP (Xia & Aggarwal, 2013)	-	-	-	89.3
3	HOJ3D(Xia *et al.*, 2012)	72.9	85.5	63.5	79.0
	EigenJoints(Yang & Tian, 2012)	74.5	76.1	96.4	82.3
	Actionlets Ensemble (Wang *et al.*, 2012b)	-	-	-	88.2
	HOD (Gowayyed *et al.*, 2013)	92.4	90.2	91.4	91.3
	(Vemulapalli *et al.*, 2014)	95.3	83.8	98.2	92.5
	(Devanne *et al.*, 2015)	-	-	-	92.1
	LM^{3}TL(Yang *et al.*, 2017b)	-	-	-	90.53
	MIMTL(Yang *et al.*, 2017a)	-	-	-	93.6
	(Jalal *et al.*, 2017)	90.8	93.4	95.7	93.3
	(Lillo *et al.*, 2017)	94.3	92.9	99.1	95.4
	3DMTG(Liu *et al.*, 2016a)	92.4	93.8	97.1	94.4
	GBSW	**93.4**	**94.9**	**98.4**	**95.6**
Procotol from (Zanfir *et al.*, 2013) (samples of subject 1, 2, 3, 4, 5 as training)					
	HON4D (Oreifej & Liu, 2013)	-	-	-	88.9
	Pose set(Wang *et al.*, 2013)	-	-	-	90.2
	Moving Pose (Zanfir *et al.*, 2013)	-	-	-	91.3
	(Lillo *et al.*, 2017)	-	-	-	93.0
	3DMTG(Liu *et al.*, 2016a)	87.50	95.8	94.7	92.7
	GBSW	**88.9**	**96.2**	**95.5**	**93.5**

accuracy, which improved the performance of (Seidenari *et al.*, 2013) and (Vemulapalli *et al.*, 2014) by 11.6% and 2.7%, respectively.

On both *MSR-Action3D* dataset and *Florence3D-Action* dataset, the proposed method

Table 3.6: Average accuracy (%) of *Cross Subject Test* on the *Florence3D-Action*.

Multi-Part Bag-of-Poses (Seidenari *et al.*, 2013)	82.0
Full skeleton(Devanne *et al.*, 2015)	85.9
Body part(Devanne *et al.*, 2015)	87.0
(Vemulapalli *et al.*, 2014)	90.9
3DMTG(Liu *et al.*, 2016a)	91.3
GBSW	**93.6**

achieved better recognition performance over 3DMTG which calculates the bag of semantic words without enhancing the discriminative motion information. The improved accuracies indicate that the discriminative information weighted using the kernel function can help the proposed method improve the ability to distinguish different activity classes.

3.4 Summary

This chapter presents the GBSW representation for human action recognition. The proposed BSW, which highlights the discriminative moving trend of each activity category via a kernel-based dynamic encoding algorithm, was aggregated with the G feature in GBSW for human action recognition. Experimental results on two public datasets have proved the compelling recognition results of the proposed approach. This outperforming performance is owed to the semantic representation and the complementary effect of the aggregation of different types of features.

Chapter 4

Moving Similarity for Human Interaction Recognition

4.1 Introduction

As an communication element, human-human interaction plays an important role in our daily life. Although the development of RGB-D sensors has motivated considerable work conducted for human action recognition, the research for human interaction is relatively unexplored. Unlike single person actions, human interaction is a behavior performed by more than one person, where the interaction relationship between people is of vital importance. Moreover, human interaction has large feature dimensions which consist of individual information as well as mutual relations. The mutual relations are typically represented by the distance between body parts in most of the existing methods (Ji *et al.*, 2014, 2015; Yun *et al.*, 2012). For example, (Ji *et al.*, 2015) associated the distance and motion features from single body part and interactive body part pairs for interaction representation. The distance property could provide useful geometric information, however, it might be not effective enough to mine intrinsic characteristics embedded in diverse interaction classes. Thus, exploring high level or semantic information could help to enhance the performance of the traditional feature representation for human interaction recognition (Ni *et al.*, 2013).

In the previous chapter, the bag of semantic moving words of each joint is constructed to represent the moving trend of skeleton joints. This feature could describe the specific tendency of skeleton joints in 3D space and has been proven to be competitive in human action recognition. Based on this feature extracted from individuals, this chapter proposes the moving similarity (MS) between body parts to describe the mutual relationship for human interaction recognition. Although several RGB-D based single person action datasets such as *MSR Action3D* (Li *et al.*, 2010) and *MSRDailyActivity3D* (Wang *et al.*, 2012b) have been collected for human action recognition, there are few RGB-D based human interaction datasets specially designed for human interaction recognition. Thus, this chapter introduces a new large RGB-D based human-human interaction dataset, namely Online Human Interaction (OHI) Dataset.

The remainder of this chapter is organized as follows: Section 4.2 introduces the proposed human interaction representation. Section 4.3 presents the collected OHI dataset. Section 4.4 reports experimental results as well as the comparison with the state-of-the-art methods. Section 4.5 summarizes the work of this chapter.

4.2 Human Interaction Representation

4.2.1 Notation of Human Body Parts

Human bodies can be divided into five parts, i.e., the left/right arm, the left/right leg and the torso. Each body part with four joints is described as following:

$$P = \{b_1, b_2, b_3, b_4\} \tag{4.1}$$

where b_i is the joint belongs to the body part, and it consists of three coordinates: $b_i = \{x_i, y_i, z_i\}$. Human-human interaction actually can be seen as the interaction between human body parts, thus the mutual relationship between humans could be described by the spatial and motion relationship among body parts in the proposed method.

4.2.2 Spatial Relationship

The distance between two interactive persons at each frame is an important spatial cue of interactions. This mutual spatial relationship between different body parts changes over the time as well as in different interaction category. For example, in the *handshaking*, the distance between the hand joints of two persons is narrowing over the sequence until the shaking is finished; while in *kicking*, the distance between the foot and the hip area narrows firstly when one person is attacking, and it begins to broad when another person is defending. This spatial relationship between joints could be effectively described by the frame-level change of Euclidean distance:

$$Dist_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2} \tag{4.2}$$

Therefore, the spatial relationship between two body parts could be represented by combining the relationship of their corresponding joints:

$$R\left(P_{p1}, P_{p2}\right) = \{Dist_{12}, ..., Dist_{ij}, ...\} \tag{4.3}$$

where $p1$ ans $p2$ are body parts, i and j are the joint from the same or different body part, which depict the intra spatial configuration of the body part itself and the inter spatial configuration of body part pairs. As shown in Fig. 4.1, the geometric relation between joints from the same body part indicates intra-relationship (the cool lines) and the geometric relation between joints from different body parts indicates inter-relationship (the pink lines).

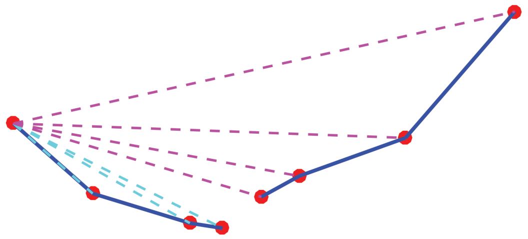


Fig. 4.1: Spatial relationship between two shaking hands.

4.2.3 Semantic Moving Similarity of Body Parts

There are various ways for performing an activity but the the moving trend of each category of acitvities is particular and represents certain meaning. Taking 'handshaking' for instance, both right hands stretch in the leftforward direction and then shake in the up and down directions. Therefore, the semantic moving direction implements the inherent meaning to the spatial relationship and thus could make features more distinctive between different interactions.

The moving direction between two consecutive frames is quantified into defined semantic words. Thus the BSW of each joint could be calculated, as shown in Fig. 4.2a, and the value in each bin quantifies the motion degree in the corresponding moving word. In Fig. 4.2b, the first row indicates the two active or inactive body parts during shaking hands. The blue histograms in the second row are the cumulative moving directions of joints (black ones) from each person, and the yellow part in the third row is the result of intersection, which reflect the degree of similarity between two body parts. Since each body part consists of four joints, its moving trend feature could be summarized by traversing its corresponding four joints:

$$\{BSW_1^p, BSW_2^p, ..., BSW_4^p\}, \tag{4.4}$$

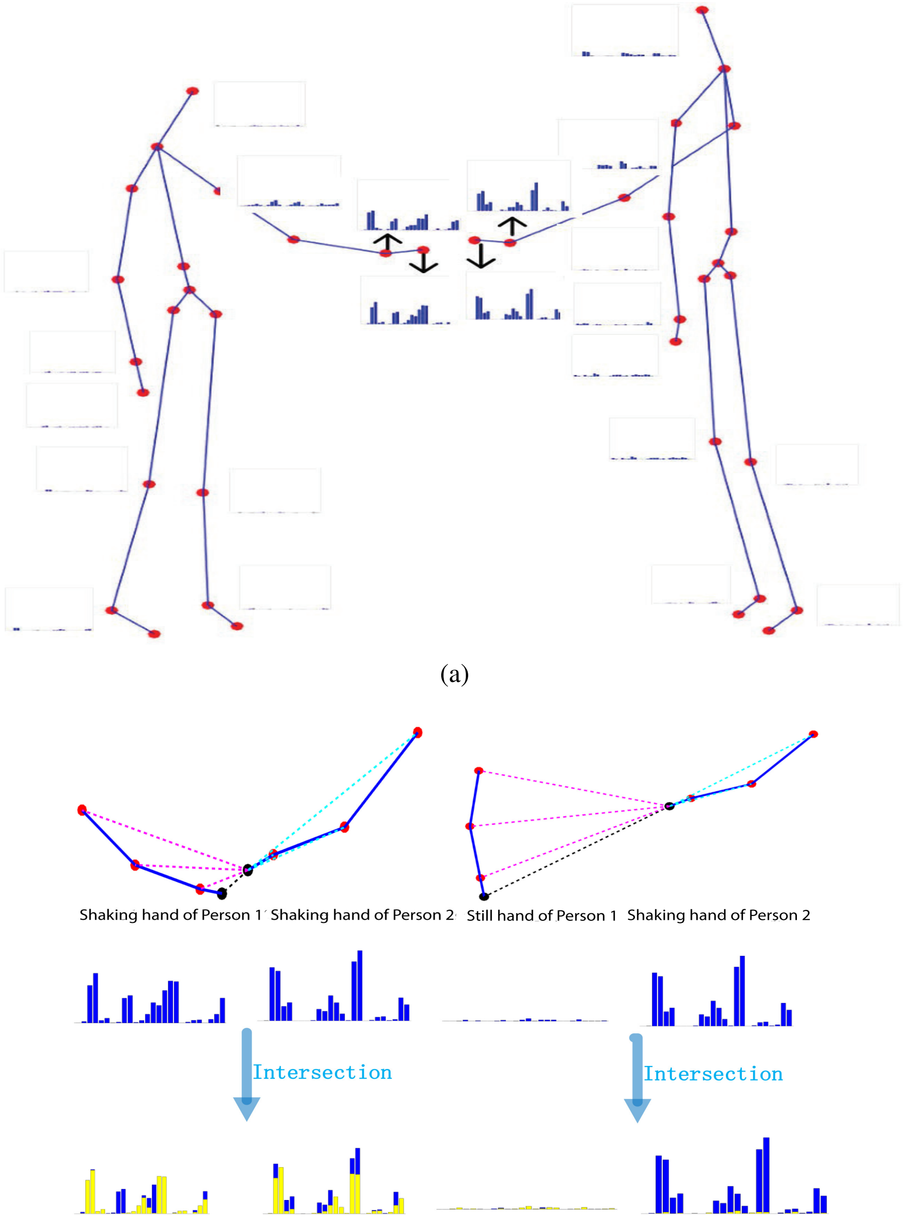
where p denotes the body part.

Histogram intersection (Swain & Ballard, 1991) was proposed for color indexing in object recognition and it can measure the degree of similarity between two histograms (Barla *et al.*, 2002). Here, the histogram intersection is adopted to count the times of direction word in one joint that have corresponding times of the same direction word in another joint. That is to say, the similarity between body parts is calculated by intersecting the moving trend histograms. The similarity between the corresponding word w from BSW$_i$ (joint i) and from BSW$_j$ (joint j) is denoted as follows:

$$\begin{aligned} &SoJ\left(BSW(\mathbf{v}_w^i), BSW(\mathbf{v}_w^j)\right) \\ &\qquad = min\left(|BSW(\mathbf{v}_w^i)|, |BSW(\mathbf{v}_w^j)|\right) \end{aligned} \tag{4.5}$$

The histogram of semantic moving words is interpolated into the same number of frames (N). By doing this, each bin in BSW having the same dimension. Thus, the revised BSW with an $N \times n$-dimensional vector could be defined as follows:

$$\widehat{BSW} = (\overbrace{1, ..., 1}^{BSW(\mathbf{v}_1)}, \underbrace{0, ..., 0}_{N-BSW(\mathbf{v}_1)}, ..., \overbrace{1, ..., 1}^{BSW(\mathbf{v}_w)}, \underbrace{0, ..., 0}_{N-BSW(\mathbf{v}_w)}) \tag{4.6}$$



(a)

(b)

Fig. 4.2: (a) Quantization of moving directions in the space for each joint. (b) Semantic similarity between body parts by histogram intersection.

Following Eq.4.5, the similarity between body parts is denoted as follows:

$$\begin{aligned} SoP_{type} &= \sum_{p=1}^{8}\sum_{q=1}^{8}\sum_{i=1}^{n}\sum_{j=1}^{n} SoJ\left(BSW_i^p, BSW_j^q\right) \\ &= \sum_{p=1}^{8}\sum_{q=1}^{8} \widehat{BSW}^p \cdot \widehat{BSW}^q \end{aligned} \tag{4.7}$$

where p and q are body parts, i and j are joints, and SoP_{type} could be SoP_{intra} or SoP_{inter}, which means intra-similarity or inter-similarity, respectively. The relationship between body parts from the same person and from two persons, respectively. $\widehat{BSW}^p$ and $\widehat{BSW}^q$ are the histogram concatenation of joints from the body part p and q, respectively. The final moving similarity (MS) of body parts for each sequence is the concatenation of all body part pairs:

$$MS = \{SoP_{intra1}, SoP_{inter1}, ...\} \tag{4.8}$$

4.2.4 Human Interaction Descriptor

Fig. 4.3 provides the framework of the proposed human interaction descriptor. The obtained skeleton joints sequences are firstly pre-processed by translating them to the body-center coordinate system as presented in Chapter 2. By doing this, the following extracted features are invariant to different locations and viewpoints. Then the intra- and inter-relationship between skeleton joints and the moving similarity among body parts are calculated to describe the spatial and motion characteristics of human interactions, respectively. By combining these two features, the SRMS descriptor is constructed as input of a SVM classifier for human interaction recognition.

4.3 Online Human Interaction Dataset

This section describes the OHI dataset. This dataset contains 23 pairs of participants with various clothing color and body size. It has 10 human-human interaction categories: *shaking hands, high waving, kicking, punching, pushing, hugging, high-fiving, approaching, departing and exchanging objects*. Fig. 4.4 shows some examples of different interaction categories. Each category is repeated for three times and some of the

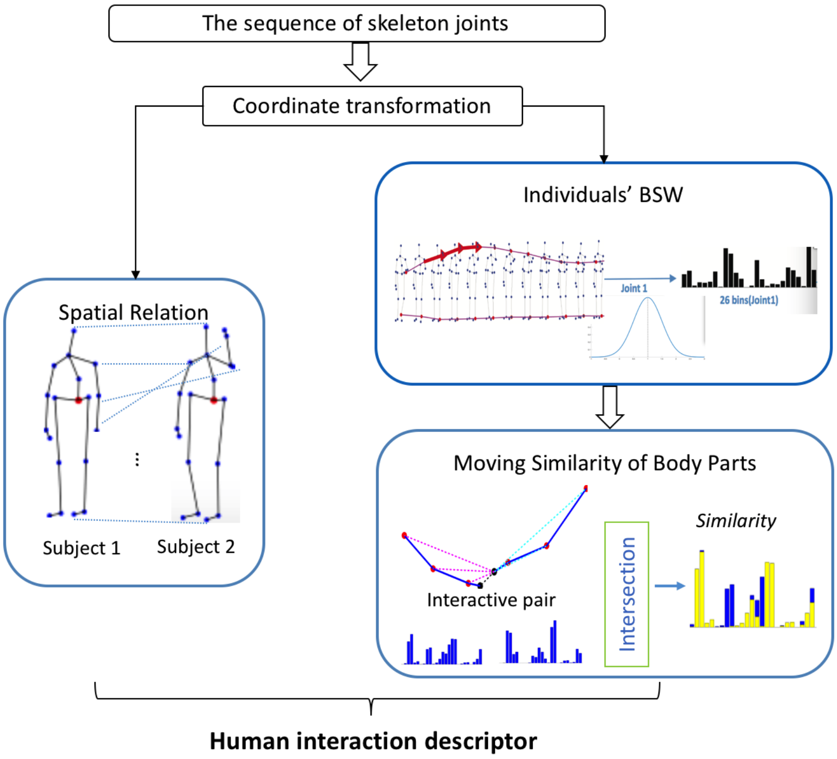
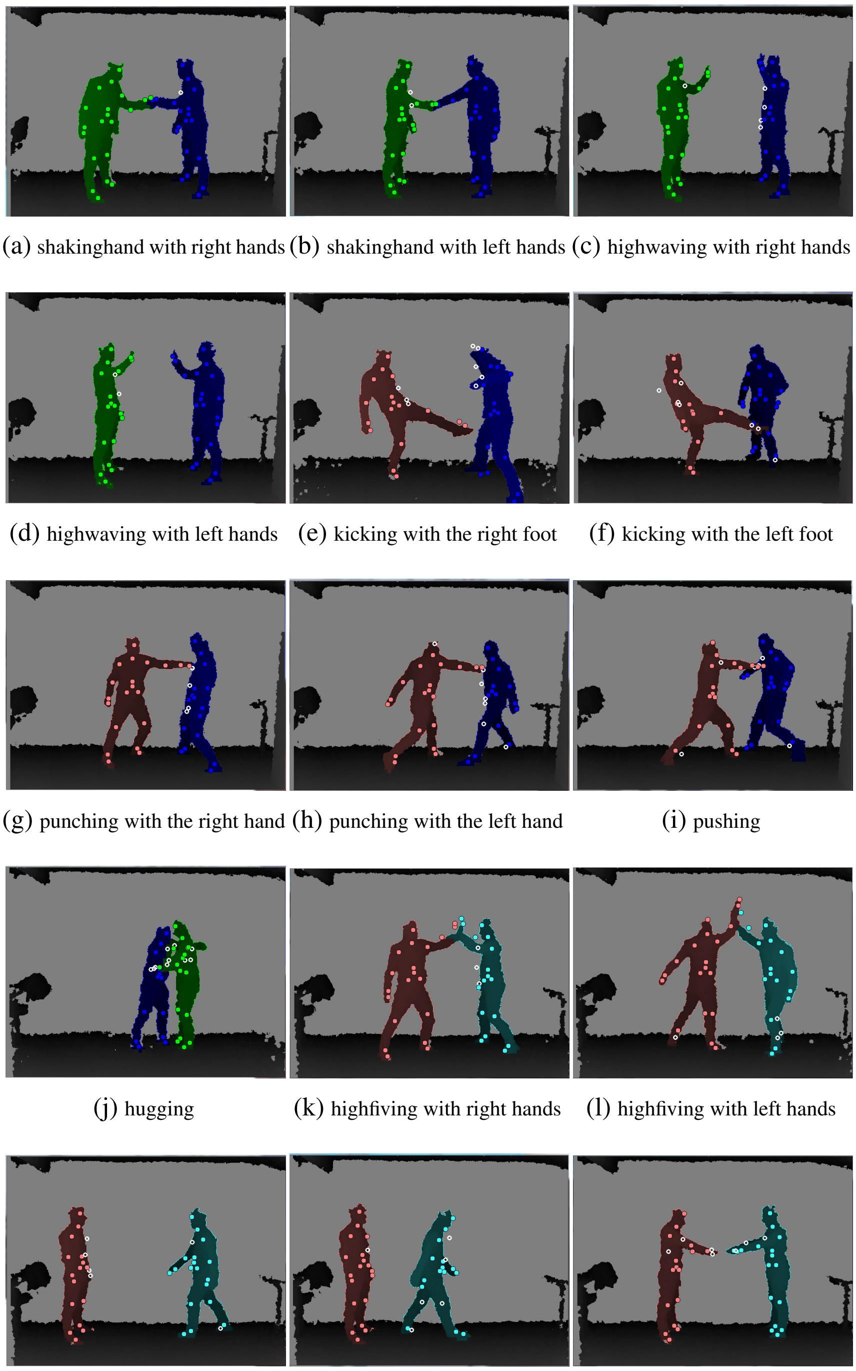

The sequence of skeleton joints
Coordinate transformation
Individuals' BSW
Joint 1
26 bins(Joint1)
Spatial Relation
Subject 1
Subject 2
Moving Similarity of Body Parts
Interactive pair
Intersection
Similarity
Human interaction descriptor

Fig. 4.3: The framework of the proposed human interaction descriptor.

categories might have more instances due to the consideration of different performing styles (the right and left side). Thus, the total number of samples is 875.

The dataset is collected by using Kinect version 1 sensor. The recored data contains RGB data, original depth data, registered depth data and skeleton data. The registered depth data to its corresponding RGB image is further provided, which is useful for motion recognition when RGB and depth are jointly used in pixel level. The resolution of RGB and depth data is 640x480 and the dataset also provides 3D coordinates of 20 skeleton joints for each subject.

There are two parts in this database. In Part I, interactions are divided into isolated sequences according to interaction categories. This part is mainly designed for the offline activity recognition. For the evaluation of online activity recognition methods, Part II is collected where each video sequence contains 10 human interactions continuously performed by one pair of subjects. During the interval of two activities, the subjects are free to perform any actions instead of standing still, which makes this



(a) shakinghand with right hands (b) shakinghand with left hands (c) highwaving with right hands

(d) highwaving with left hands (e) kicking with the right foot (f) kicking with the left foot

(g) punching with the right hand (h) punching with the left hand (i) pushing

(j) hugging (k) highfiving with right hands (l) highfiving with left hands

(m) approching (n) departing (o) exchanging

Fig. 4.4: Interaction samples of depth images and skeleton joints on OHI dataset

database closer to practical scenarios as well as challenging. Table 4.1 lists the comparison of the OHI dataset with the existing interaction datasets. Compared to the existing datasets, the OHI dataset has four advantages:

Table 4.1: RGB-D sensor based human interaction datasets. HHI: human-human interaction.

Dataset	Interactions	Participants	Samples	Data types	Video type

SBU Kinect
Interaction
Yun *et al.* (2012)
	8 HHI	7 subjects
		21 sets
			300	RGB(640x480)
				depth(640x480)
				skeleton(15 joints)
					Trimmed

K3HI Hu *et al.* (2013)	6 HHI	15 subjects	320	skeleton(15 joints)	Trimmed

ShakeFive
van Gemeren *et al.* (2014)
	2 HHI	37 subjects	100	RGB(640x480)
				skeleton(20 joints)
					Trimmed

G3Di
Bloom *et al.* (2014)
	6 Virtual
	interactions
		12 subjects
		6 sets
			-	RGB
				depth(640x480)
				skeleton(20 joints)
					Trimmed

ShakeFive2 van Gemeren *et al.* (2016)	8 HHI	-	153	skeleton	Trimmed

ISR-UoL
Coppola *et al.* (2016)
	8 HHI	6 subjects
		sets
			-	RGB(24 bits)
				depth(8,16 bit)
				skeleton(15 joints)
					Trimmed

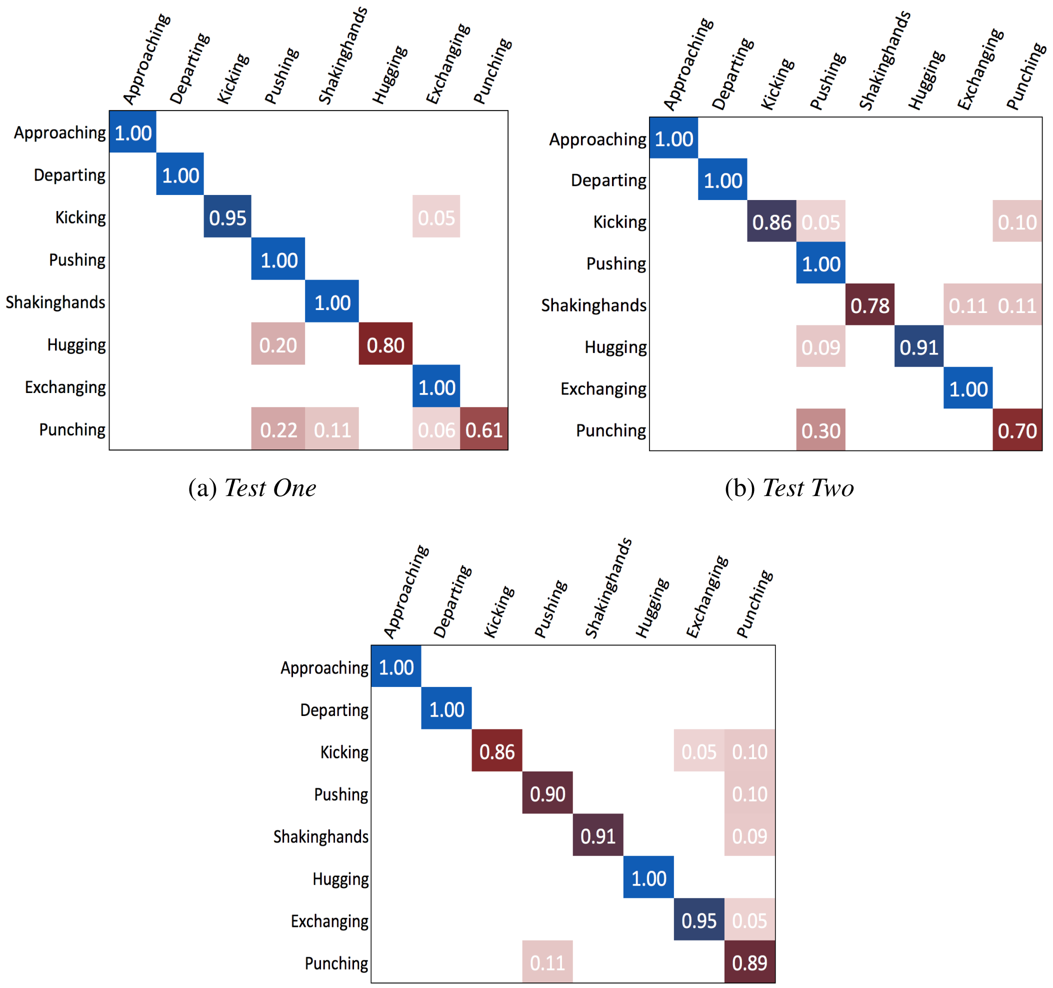
Online Human Interaction	10 HHI	13 subjects 23 pairs	900	RGB (640x480) , Depth (640x480) , Registered depth, skeleton (20 joints) , 30 fps	Trimmed Continuous

1. More interaction samples: this dataset has around 900 interactions, which is 3 times than that of (Yun *et al.*, 2012);

2. More complex: the performing habit of actors is considered by performing either the right or left side;

3. The registered depth image: the value in depth maps is registered to the corresponding RGB images. The registered depth information is useful for the segment of human body in RGB images and also provides convenience for jointly using RGB and depth data in pixel level;

4. Extra continuous activity videos are provided for the research of online performance. The time series between two activities (subjects are performing as they want) is considered as the neutral activities. Different from the neutral activities in (Huang *et al.*, 2014) where subjects do not have actions just standing in the most cases, the OHI database is closer to the reality and more challenging due to the larger variability of the neutral activities.



(a) *Test One*

(b) *Test Two*

(c) *Cross Subjects Test*

Fig. 4.5: Confusion Matixes on *SBU Interaction dataset*

4.4 Experimental Evaluation

As mentioned before, there is few publicly available human interaction datasets and most of the researchers evaluated their algorithms on the SBU Interaction Dataset (Yun *et al.*, 2012). Therefore, the proposed human interaction approach is tested on the SBU Interaction Dataset (Yun *et al.*, 2012). Owing to the newly collected OHI dataset, some experiments are conducted on this dataset to demonstrate the performance of the proposed approach. The test results on the SBU dataset is compared with the state-of-the-art approaches.This dataset contains examples of eight different interaction classes: *approaching, departing, kicking, punching, pushing, hugging, shaking hands, exchanging something*. All the videos were collected in the same laboratory environment from a third-person perspective. The majority of the interactions involve acting-reacting relation. The testing on this dataset contains the *Test One*, *Test Two* and *Cross Subject Test*.

4.4.1 Results on SBU Interaction Dataset

Having computed above feature descriptors, a linear SVM (Chang & Lin, 2011) algorithm is then applied for human interaction classification. To test the recognition ability of the proposed human interaction descriptor, the confusion matrices that indicate the confusion among activity categories and the comparison to the state-of-the- art are conducted on the *SBU Interaction dataset*. In Fig. 4.5, it can be seen that the proposed method is able to successfully classify *approaching* and *departing* in *Test One*, *Test Two*, and *Cross Subject Test*. The most common confusion is between *pushing* and *punching* in all tests due to their similar poses.

Table 4.2 reports the recognition results of listed methods.

It shows that the recognition rates of the proposed algorithm on different tests are over 90%. The average rate 92.75% is 2.25% higher than the best performance of listed skeleton-based methods reported in (Baradel *et al.*, 2017), meaning that the correlation feature explored in the proposed method could extract high-level information from the movement of skeleton joints, thus helps to reinforce the performance of discriminating complex human interactions. It should be noted that the results of RHI and the improved method in (Baradel *et al.*, 2017) outperform that of the proposed method due to the combination of RGB data and skeleton joints.

Table 4.2: Recognition Accuracy (%) on *SBU Interaction dataset*.

State-of-the-art	Velocity features (Yun *et al.*, 2012)	48.4
	Plane features (Yun *et al.*, 2012)	73.8
	Joint features (Yun *et al.*, 2012)	80.3
	(Ji *et al.*, 2014)	86.9
	CFDM (Ji *et al.*, 2015)	89.4
	CHARM (Li *et al.*, 2015)	83.9
	HBRNN (Du *et al.*, 2015b)	80.35
	RHI (Gori *et al.*, 2015)	93.08
	Co-occurrence LSTM (Zhu *et al.*, 2016d)	90.4
	(Song *et al.*, 2017)	91.51
	Skeleton (Baradel *et al.*, 2017)	90.5
	Skeleton+RGB (Baradel *et al.*, 2017)	94.1
Proposed	*Test One*	**92.75**
	Test Two	**91.67**
	Cross Subjects Test	**93.84**
	Average	**92.75**

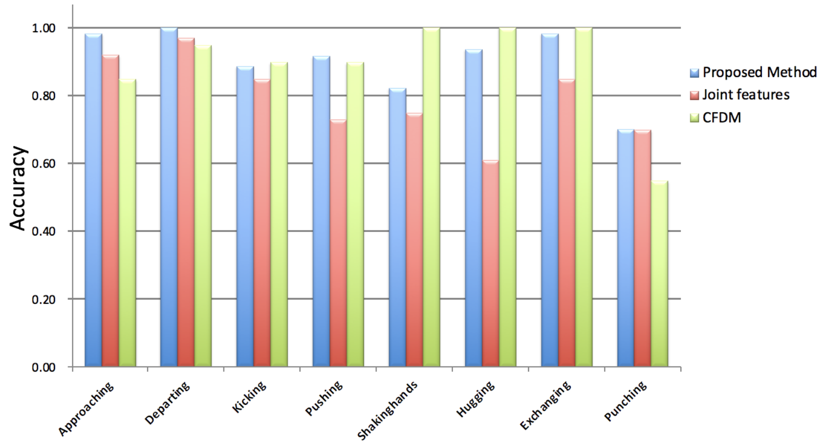


Fig. 4.6: Comparison of CFDM, Joint feature and Proposed Method by categories on SBU interaction dataset.

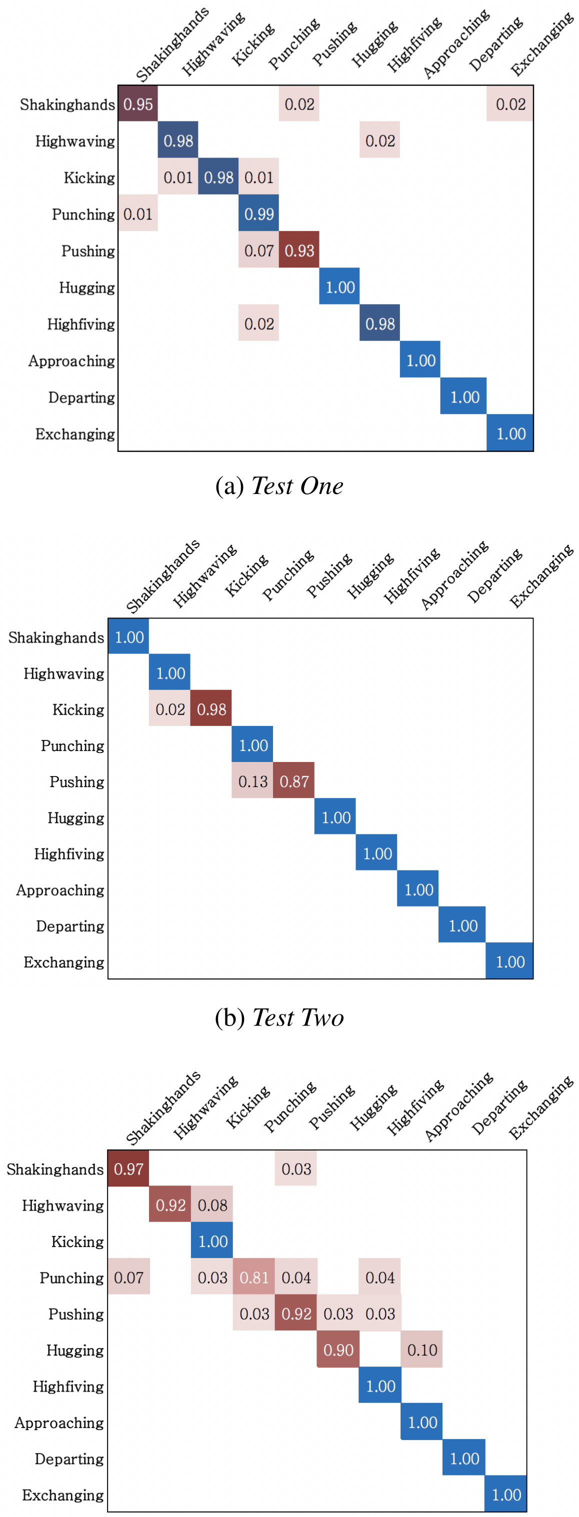
Fig. 4.6 gives the detailed recognition accuracy comparison of each category among Joint features (Yun *et al.*, 2012), CFDM (Ji *et al.*, 2015), and the proposed method. Compared to Joint Features in (Yun *et al.*, 2012), the proposed method achieves better recognition on most of the interactions, especially on *punching*, *hugging* and *exchanging*. Furthermore, the accuracies of most categories are higher than CFDM, apart from *shakinghands, hugging* and *exchanging*.

4.4.2 Experiment on OHI Dataset

For the newly collected dataset, the *Test One*, *Test Two*, and *Cross Subjects Test* setting are used for evaluation. Fig. 4.7 shows the confusion matrices. With the proposed method, the recognition rates in most interaction categories are over 90%, and some reach to 100%. Although the similarity between interactions like *pushing* and *punching* is huge, the rates that *punching* is unexpectedly recognized as *pushing* and *pushing* is unexpectedly recognized as *punching* are slow (0.03 and 0.04, respectively). Because the motion trend in the early stage of *hugging* is similar with that of *approaching*, the possibility of *hugging* recognized as *approaching* is relatively high (0.10) in *Cross Subjects Test*.

4.5 Summary

In this chapter, an effective human interaction representation is proposed. Firstly, the geometric information among skeleton joints is calculated to represent the temporal evolution of the mutual spatial relationship between interactive persons. Then, based on the bag of semantic moving words, the moving similarity between body parts is extracted to represent the motion relationship between interactive subjects. Finally, both the spatial and motion features are combined for human interaction recognition. The outstanding performance (e.g., 92.75% on SBU Interaction) indicates that the spatial relationship and moving similarity explored in the proposed method can effectively describe the mutual relationship of interactive body parts, thus helps to reinforce the performance of discriminating different human interactions. In addition, the challenging OHI dataset was introduced to be served for the evaluation of human interaction recognition methods.



(a) *Test One*

(b) *Test Two*

(c) *Cross Subject Test*

Fig. 4.7: Confusion Matixes on OHI dataset

Chapter 5

Skeleton Motion Distribution based Activity Detection

5.1 Introduction

Although significant work has been done for offline action recognition where the actions are pre-segmented by providing the starting and ending frames (Du *et al.*, 2015b; Guo *et al.*, 2017; Jalal *et al.*, 2017; Kong & Fu, 2015b; Shahroudy *et al.*, 2016c), their performance remains unclear when applied to realistic scenarios where no prior information regarding the action's trigger time is available. Most of the common situations require the algorithms to automatically process the data stream without any prior information (De Geest *et al.*, 2016). That is to say, an online system is required to recognize actions from untrimmed videos by answering 'when does the action happen?' and 'what action is happening?' (De Geest *et al.*, 2016), as shown in Fig. 5.1.

Some work has been done for the similar task named early action recognition (Hoai & De la Torre, 2014; Kong *et al.*, 2014; Ryoo, 2011), which predicts actions before they are fully finished. However, this type of methods assumed that the starting time of actions is known beforehand. This solution can only be regarded as a partial answer to online action recognition since more attention is focused on action classification instead of detection. Compared to isolated action recognition and early action recognition, online action recognition is significantly more challenging for two reasons: firstly, the boundaries of various action categories need to be detected accurately; secondly,

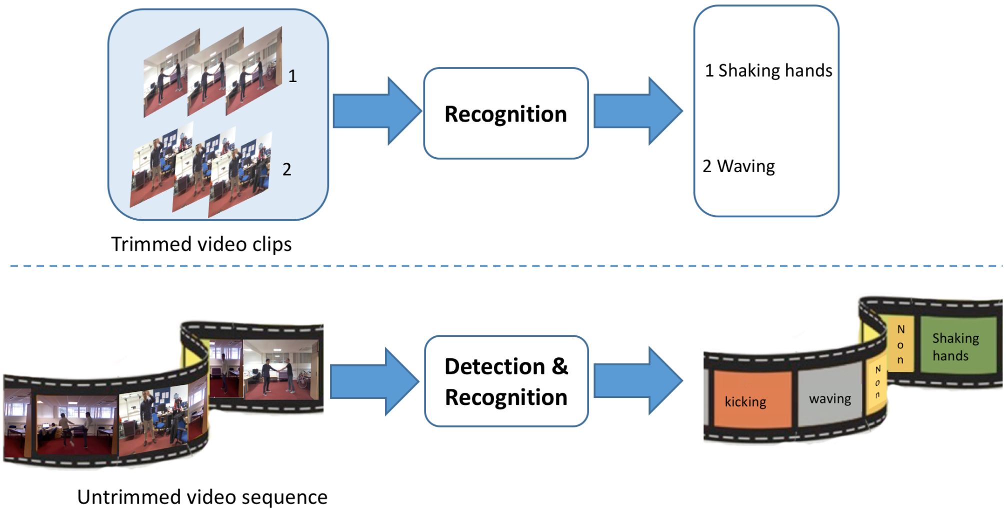


Fig. 5.1: The comparison between isolated activity recognition and online activity recognition.

only partial actions might be available for recognition due to the performance of action detection algorithm, thus the action recognition algorithm should be capable of recognizing actions from different action fragments.

This chapter proposes a skeleton motion distribution based method to simultaneously perform action detection and classification in continuous videos. The unique movement characteristics of each action class make the corresponding motion sequence have different distribution from each other. The human motion is modeled as a stochastic distribution using kernel density estimation, which has witnessed great success in detecting outliers (Latecki *et al.*, 2007) and background subtraction (Elgammal *et al.*, 2002; Mittal & Paragios, 2004). Since the change from one action to another results in the distribution property at the beginning of the new action deviating so much from the previous action in a video sequence. Therefore, the occurrence time of actions could be detected depending on this change. Once an action is detected, a snippet-based classifier is designed to process the observed video to achieve action classification. This classifier is performed in a fragment level which could reduce the influence of false detections caused by noises.

The remainder of this chapter is organized as follows: Section 5.2 introduces the proposed data pre-processing method. Section 5.3 introduces the proposed action detection and recognition method. Section 5.4 reports experimental results as well as the comparison with the state-of-the-art methods. Section 5.5 summarizes the work of this chapter.

5.2 Data Pre-processing

The offset displacement of each joint is utilized to describe skeleton motion. Given a human skeleton motion sequence $d_1, d_2, ..., d_n \in N$, where N is the frame number. The coordinate of joints is translated to the hip-center coordinate system to make the feature invariant to different locations. $d_t = [\Delta^1_{x_t,y_t,z_t}, \Delta^2_{x_t,y_t,z_t}, ..., \Delta^{20}_{x_t,y_t,z_t}]$ is the offset displacement feature of skeleton joints in 3D space at time t. Herein, skeleton displacement offset is computed as follows:

$$\begin{cases} \triangle x^i_t = x^i_t - x^i_1, \\ \triangle y^i_t = y^i_t - y^i_1, \\ \triangle z^i_t = z^i_t - z^i_1, \end{cases} \tag{5.1}$$

x^i, y^i, z^i are coordinates of the $i - th$ joint. $\triangle x^i_t, \triangle y^i_t, \triangle z^i_t$ are displacement offset in three directions with respect to the initial position. To obtain a compact representation, it is assumed that the skeleton motion feature lies on a low dimensional manifold embedded in the ambient space, thus locality preserving projection (LPP) (He & Niyogi, 2003) is applied to reduce the dimensions of skeleton motion d. A transformation matrix W is calculated to map motion data d to a set of points p in a low dimensional space, such that $p = W^T d$. LPP can optimally preserve the local neighborhood structure of the data while mapping it to a lower dimensional space. Following the procedure listed in Table. 5.1, the skeleton motion is represented in a subspace by a linear transformation $p = W^T d$.

5.3 Skeleton Motion Distribution

The density distribution of two actions' skeleton motion is obtained using adaptive kernel density estimation (Latecki *et al.*, 2007), which is capable of estimating the

Table 5.1: The procedure of representing human skeleton motion in low dimensional space using LPP algorithm.

```
Input: human skeleton motion set $D = d_1, d_2, ..., d_t$ at $t = 1, 2, ..., T$.
Output: mapped representation on low dimension space $p = W^T d$
Step 1: Constructing the adjacency graph.
        Let $G$ denote a graph with $m$ nodes.
        An edge is put between nodes $i$ and $j$ if data $d_i$ and $d_j$ are 'close'.
        LPP will choose the projections which can optimally preserve this
        adjacency graph.
Step 2: Choosing the weights.
        $W$ is a symmetric matrix with weight value $w_{ij} = 1$
        if and only if nodes $i$ and $j$ are connected, otherwise $w_{ij} = 0$.
Step 3: Computing Eigenmaps.
        The objective function: $min \sum_{i,j}^{n} ||p_i - p_j||^2 \boldsymbol{S}$;
        According to $p = W^T d$,
            $\frac{1}{2} \sum_{i,j}^{n} (p_i - p_j)^2 \boldsymbol{S} = \frac{1}{2} \sum_{i,j}^{n} (W^T d_i - W^T d_j)^2 \boldsymbol{S}$
            $= W^T D(D - S) D^T W = W^T DLD^T W$
        Compute the eigenvectors and eigenvalues for the generalized
        eigenvector problem: $DLD^T W = \lambda XCX^T W$
Step 4: Mapping data. $p = W^T d$
```

probability density of data samples without any assumptions of underlying data distribution. Assuming the mapped motion data $\{p_1, p_2, ..., p_n\}$ is the random variable in a feature space from a distribution with an unknown density $q(p)$. The estimated $\widehat{q}(p)$ at point p is obtained using the following kernel density estimation function:

$$\widehat{q}(p) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(p_i)^k} K(\frac{p - p_i}{h(p_i)}) \tag{5.2}$$

where k is the dimensionality of data samples, and $h(p_i)$ is the bandwidth at point p_i. $K(\cdot)$ is a kernel function. Herein, a multivariate Gaussian function with zero mean and unit standard deviation is adopted. Thus, Eq. 5.2 could be denoted as follows:

$$\widehat{q}(p) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(h(p_i))^k} \cdot \frac{1}{(\sqrt{2\pi})^k} exp(-\frac{||p - p_i||^2}{2(h(p_i))^2}) \tag{5.3}$$

Motion data is mostly multi-modal, even in the same action category, resulting in that the data from different modality has different density. The estimated density might be not precise if the whole set of motion data is used to estimate the density. On the other hand, the distribution of a point should have the similar distribution with its neighbor points. To this end, the density estimation of a point is adapted by using its neighbor $Km(p)$. Thus, Eq. 5.3 is modified as follows:

$$\widehat{q}(p) \propto \frac{1}{m} \sum_{p_i \in Km(p)} \frac{1}{(\sqrt{2\pi}h(p_i))^k} exp(-\frac{||p-p_i||^2}{2(h(p_i))^2}) \tag{5.4}$$

where $Km(p)$ includes m samples ($m << n$) belonging to the neighbor of the point p. Compared to the holistic comparison, the local measure depending on $Km(p)$ yields more effective density estimation by reducing computation costs from the whole data set (n samples) to local neighbors (m samples).

To enhance the robustness of the density estimation function against the change of data distribution for different subjects or action instances, an adaptive bandwidth, $h(p_i) = h \cdot d_m(p_i)$, is achieved by considering the distance $d_m(p_i)$ between the point p_i and its $m-th$ neighbor, as shown in Eq. 5.5. This improvement makes the bandwidth small in density action samples while big in sparse ones, thus enables the density estimation function adaptive to various data density.

$$\widehat{q}(p) \propto \frac{1}{m} \sum_{p_i \in Km(p)} \frac{1}{(\sqrt{2\pi}h \cdot d_m(p_i))^k} exp(-\frac{||p-p_i||^2}{2(h \cdot d_m(p_i))^2}) \tag{5.5}$$

Fig. 5.2 shows an example of action detection, where the red dots and the blue stars are samples from the normal action and *running* action, respectively. The blue star in the rectangle is the starting point of *running*. The estimated density of *running* points even at the beginning of the action has huge difference compared to that of the normal action, which means the density distribution of a different action starts to deviate significantly from the previous action at its beginning. Therefore, the relative density relationship is used to effectively describe this difference, as denoted in Eq. 5.6:

$$LDR(p) \propto \frac{\sum_{p_i \in Km(p)} \frac{\widehat{q}(p_i)}{m}}{\widehat{q}(p) + c \cdot \sum_{p_i \in Km(p)} \frac{\widehat{q}(p_i)}{m}} \tag{5.6}$$

where $LDR(p)$ denotes the ratio between the density at p and the average density of its neighbors. c is a scaling constant to avoid infinity values of LDR caused by very small estimated density at the point p.

Fig. 5.3 shows that the local estimated density of targeted actions has apparent difference compared to normal actions, and this value will convert dramatically at its

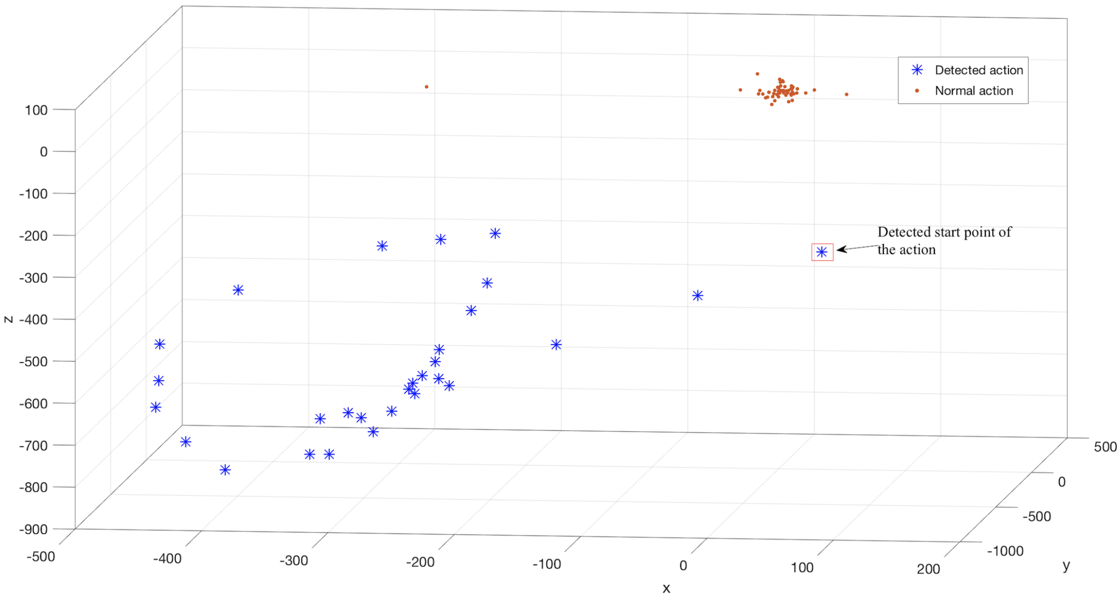


Fig. 5.2: Example of action detection.

starting and ending due to the change of the density distribution from action to action. This sudden convert referred to as action boundaries in action sequences can be

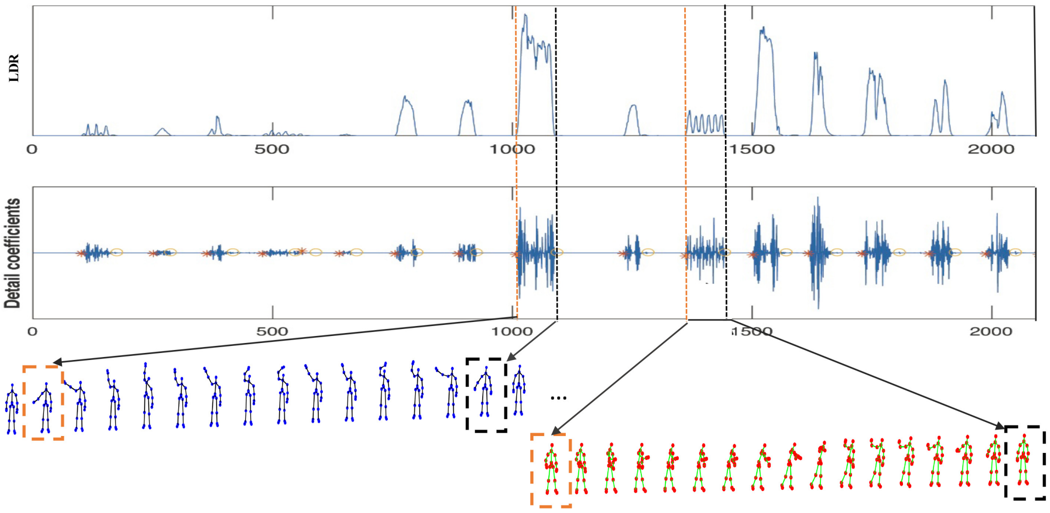


Fig. 5.3: The local density ratio of a continuous video and action starting /ending points detection through wavelet transform.

regarded as a type of impulses, which could be detected via wavelet transform. The wavelet transform is a powerful technique for analyzing irregular data, owing to its great capacity in providing the frequency and corresponding time location information of signals. This makes the wavelet transform suitable for detecting impulses occurring at any time. Approximation and detail coefficients are outputs from the low-pass and high-pass filter, respectively. The significant difference in density distribution between the occurrence of the action and its neighbors allows the detection of impulses during a time series with a high accuracy.

5.3.1 Snippet-based Action Recognition

Action durations exhibit considerable variability and only partial action observations might be available due to the performance of action detection algorithm. To address this issue, features from snippets incorporating partial segments of actions in different performing stages are alternatively extracted. Fig. 5.4 shows a schematic illustration of the proposed classification. At training time, action snippets are randomly generated from untrimmed videos using a sliding window strategy. The proposed snippet-based classifier takes advantage of local temporal information, thus makes it robust to variations in execution time.

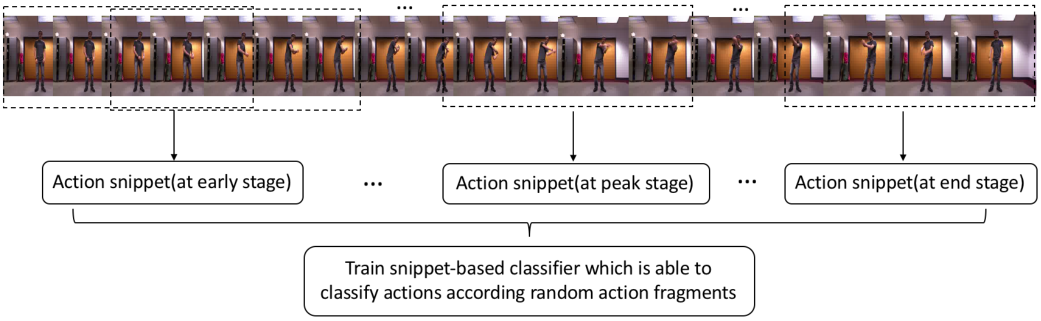

...
...
Action snippet(at early stage)
...
Action snippet(at peak stage)
...
Action snippet(at end stage)
Train snippet-based classifier which is able to classify actions according random action fragments

Fig. 5.4: Snippet-based classifier. Action snippets in different performing stage are generated from the continuous video via sliding window.

Compared to recognizing actions after they are totally completed, recognizing actions from partial action observations is more challenging due to the limited information. To make the feature descriptor discriminative for different categories, the 3D moving trend feature of joints proposed in Chapter 3 is adopted for action representation. The moving directions of skeleton joints, calculated using Eq. 5.7, are divided into various semantic words, and then a histogram which can quantitatively reflect the moving trend property is built by accumulating directions over the whole sequence.

$$\mathbf{v_t^i} = \{x_t^i - x_{t-1}^i, y_t^i - y_{t-1}^i, z_t^i - z_{t-1}^i\} \tag{5.7}$$

Cosine similarity and displacement are combined for soft voting during histogram quantification, as formulated in Eq. 5.8:

$$cos\theta_j^i(t) = \frac{\mathbf{v_j} \cdot \mathbf{v_t^i}}{\|\mathbf{v_t^i}\| \|\mathbf{v_j}\|}, j \in [1, m] \tag{5.8}$$

where $\mathbf{v_j} \in \mathbf{V}$ is the defined semantic words.

5.3.2 Framework of Online Action Recognition

Table 5.2 lists the detailed procedure of action detection and recognition. The action boundaries are detected depending on the local density relationship in continuous action sequences. And then the snippet-based classification is processed to continuously

Table 5.2: Framework of the proposed online action recognition.

```
Input: skeleton motion set $D = d_1, d_2, ..., d_t$ at $t = 1, 2, ..., T$.
Snippet-based classifier $C$, window size $length$, stride $step$ and threshold $\delta$.
Output: Start points $StartP$ and end points $EndP$ of actions, action class $Label$
Initialization: $StartP = 0$, $EndP = 0$.
While $t < T$
  · Map data $d_t$ to a low dimension space using LPP:$p_i = W^T d_i$;
  · Compute local density relationship $LDR(p_i)$;
  · Detect $StartP$ and $EndP$ according to detail coefficients $cD1$ and $\delta$ using dwt:
    $cD1[t] = \sum_{-\infty}^{+\infty} y[k]h[2t-k]$, where $h$ is the high-pass filter;
  If $cD1(t) > \delta$
    $StartP = t$;
    Start snippet-based action recognition from time $t$ using sliding window;
    Assign each snippet a specific class label $label(i)$ by the classifier $C$.
  util $cD1(t) < \delta$
    $EndP = t$;
    Smooth the labels of snippets from the detected sequence over time $StartP$
    to $EndP$;
    Select the final $Label$ with highest probability to the detected sequence;
  End
End
```

classify partial actions. The proposed method achieves lower computation cost compared to continuous recognition over the whole video, because the action recognition is processed intermittently if the occurrence of actions is detected. In addition, the classifier performing in fragment level could reduce the influence of false detections and thus improve the performance.

5.4 Experiment Results

The effectiveness of the proposed method in detecting and recognizing actions from continuous skeleton sequences is evaluated on the MAD database (Huang *et al.*, 2014). At test time, local density relationship is computed using Eq. 5.6 and then passed through the wavelet transform for detecting action boundaries. For a fair comparison, five-fold-cross-validation is performed as set in (Huang *et al.*, 2014). At train time, a snippet-based 36-class classifier is trained, where action snippets representing different action stage of one class were picked up from videos.

Table 5.3 gives the comparison of the average performance in terms of *precision percentage (Prec)* and *Recall percentage (Rec)* that defined in (Huang *et al.*, 2014). *TN*: number of correctly detected events who has 50% overlap with the ground truth event; *GTN*: Number of all ground truth events; *DN*: number of detected events. $Prec = \frac{TN}{DN}$ and $Rec = \frac{TN}{GTN}$. From the table, it can be seen that the proposed method achieves over 80% accuracies in $Prec$ and Rec which outperform the other compared methods. Specifically, the $Prec$ of the proposed method is approximately 60% higher than SVM+DP (Hoai *et al.*, 2011), 25% higher than SMMED (Huang *et al.*, 2014), and 8.7% higher than ENB (Escalante *et al.*, 2016). On the other hand, the proposed

Table 5.3: Average Detection and Recognition results on *MAD database*(%)

Methods	*Prec*	*Rec*
SVM+DP (Hoai *et al.*, 2011)	28.6	51.4
SMMED (Huang *et al.*, 2014)	59.2	57.4
ENB (Escalante *et al.*, 2016)	76.1	73.6
Method in (Devanne *et al.*, 2017a)	72.1	79.7
Proposed	**84.8**	**80.8**

method improves the Rec rate by 7.2% compared to the result of ENB. Higher precision percentage and recall percentage of the proposed method indicate that it has the capability to precisely recognize actions within detected actions and accurately detect actions from continuous videos.

The action detection performance of different methods on two sample action sequences is also compared in Fig. 5.5. It can be seen from these color bars that the

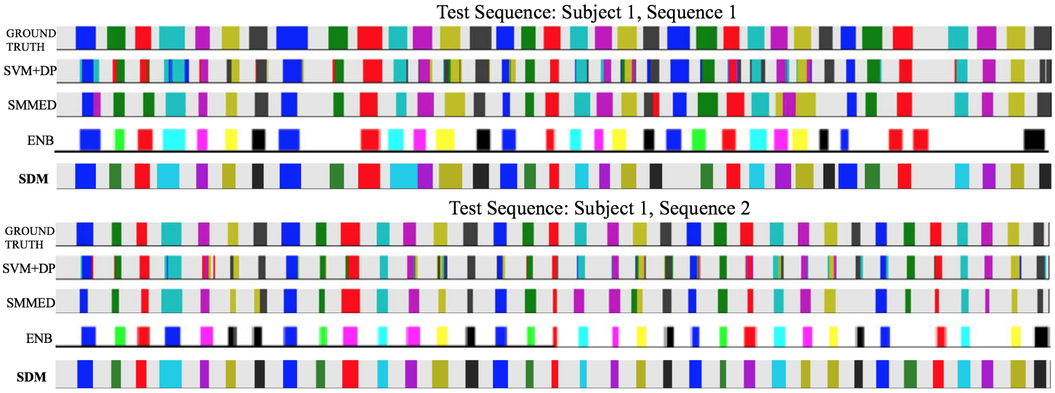


Fig. 5.5: The comparison of the proposed method with SVM+DP, SMMED, and ENB on two test sequences in the MAD database.

performance of the proposed method is the closest to the ground truth compared to other listed methods. Although SVM+DP and SMMED could detect almost all action occurrence, the accuracy of classification within a detected action is relatively low. In spite of better performance in classification within detected actions of ENB, its missing detection rate decreases. The proposed method is able to detect actions correctly with a lower missing detection rate than ENB and performs a higher classification accuracy than all listed methods. Furthermore, Fig. 5.6 reports the average detection accuracy of each action class. The majority of actions could be detected with the accuracy around 80%. The relatively low accuracy in action 4 (‘walking’), action 22 (‘Right Arm Dribble’), and action 23 (‘Right Arm Pointing to the Ceiling’) might be the result of the action similarity.

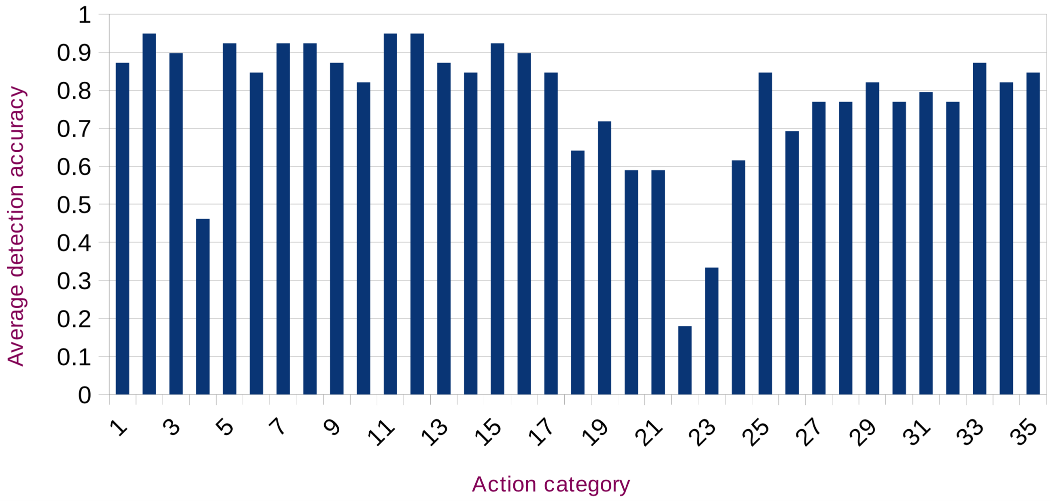


Fig. 5.6: Average detection accuracy of each action category.

5.5 Summary

This chapter proposed a skeleton motion distribution based method to deal with the action detection and recognition problem in online action recognition. An adaptive density estimation function was developed for calculating the density distribution of skeleton motion in different actions. The transition of the density distribution from action to action was investigated for effective action detection. Furthermore, a snippet-based classifier which can handle action fragments was trained for the sequential action recognition once actions were detected. Although the comparison with the state-of-the-art methods in the publicly available database has shown that the proposed method obtained better results including detection and classification performance. However, the performance of accurately identifying uncompleted actions from partial observations still remains challenge.

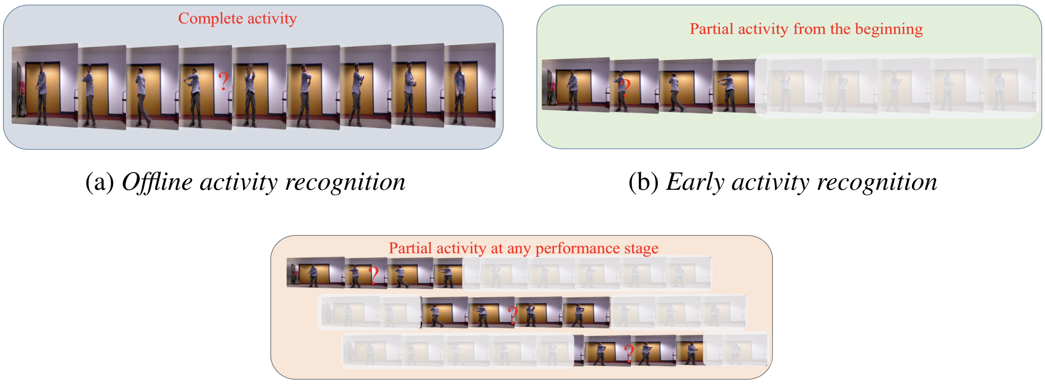
Chapter 6

Multi-stage Soft Regression for Effective Online Activity Recognition

6.1 Introduction

Traditional offline activity recognition approaches aim to recognize pre-segmented activities whose start time and end time are manually extracted (Guo *et al.*, 2017; Jalal *et al.*, 2017; Liu *et al.*, 2017d; Oreifej & Liu, 2013; Qiao *et al.*, 2017; Shi *et al.*, 2017; Wang *et al.*, 2015a; Zanfir *et al.*, 2013; Zhang *et al.*, 2018), thus the recognition results are typically given after the event occurred. However, in most of the practical scenarios, it is difficult to know the boundary of activities ahead and the recognition results need to be given during the activity period with low latency (Cai *et al.*, 2016). Thus, online activity recognition which aims to detect and recognize activities as soon as possible in a continuous video stream is significantly important in such applications. To release the constraint of manually segmenting activities in traditional activity recognition, some approaches executed activity segmentation and classification simultaneously in an input video (Evangelidis *et al.*, 2014; Kulkarni *et al.*, 2015). Although activities are segmented automatically, the classification is performed until full activities are detected.

Compared to offline activity recognition and early activity recognition, online activity recognition is more complex in that it needs to simultaneously and quickly perform activity detection and recognition in a continuous video stream (Wang *et al.*, 2018), as shown in Fig. 6.1. One of the most challenging problems in online activity


Complete activity
?
Partial activity from the beginning
?
Partial activity at any performance stage
?
?
?

(a) *Offline activity recognition*

(b) *Early activity recognition*

(c) *Online activity recognition*

Fig. 6.1: Comparison of offline activity recognition, early activity recogniton, and online activity recognition.

recognition is the partial activity observation problem that only part of the activity can be observed due to the incomplete sequence acquisition. The partial activity observation is possible at arbitrary performance stages, with a large inter-class variability and intra-class similarity, and the available information of activities is limited.

This chapter addresses the task of recognizing activities with partial activity observations by formulating it as a Multi-stage Soft Regression (MSR) problem. Multiple score functions that measure the compatibility between a video segment and an activity label are collaboratively learned in the MSR framework. Human activities are divided into three performance stages, namely *start, peak*, and *end*. At training time, segments spanning all performance stages are collected to make score functions robust and effective to arbitrary activity fragments, as shown in Fig. 6.2. Different activities may have similar observations at their starting or ending stages so that uniformly using these observations for training might result in poor recognition performance. The discriminative power of the regression model is enhanced by collaboratively learning three score functions with a focus on different stages. Furthermore, the inherent evolution of segments from adjacent performance stages is modeled by introducing a soft label strategy into the learning formulation.

The remainder of this chapter is organized as follows: Section 6.2 introduces the process of the proposed method. Section 6.3 reports various experimental results as

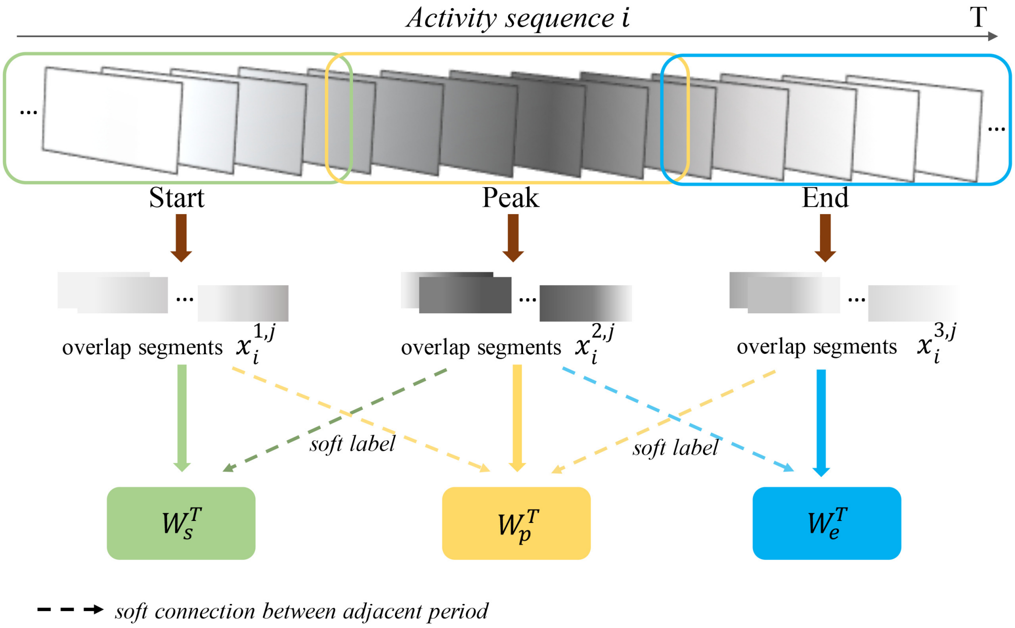

Activity sequence i
T
...
...
Start
Peak
End
...
overlap segments $x_i^{1,j}$
...
overlap segments $x_i^{2,j}$
...
overlap segments $x_i^{3,j}$
soft label
soft label
W_s^T
W_p^T
W_e^T
soft connection between adjacent period

Fig. 6.2: The overview of the proposed MSR framework.

well as the comparison with the state-of-the-art methods. Section 6.4 summarizes the work of this chapter.

6.2 Model Formulation

6.2.1 Motivation

The offline activity recognition methods suffer from a drawback that the recognition results can only be given after the action event. Observing this problem, many researchers proposed the early activity recognition to obtain the recognition results during the activity period (Cai *et al.*, 2016; Hoai & De la Torre, 2014; Huang *et al.*, 2014). However, it still relies on pre-segmented sequences by providing the starting point of the activity. These methods simplify the problem and over inflate the performance, thus lack the applicability to practical applications. To deal with the previous problems, online activity recognition that processes a continuous video stream has been appealing in recent research. The partial activity observation problem mainly caused by the incomplete sequence acquisition, makes it greatly challenging due to following reasons:

1) the partial observation is possible from an arbitrary performance stage, with a large inter-class variability and intra-class similarity; 2) activities need to be recognized as accurately as possible based on the limited information from the partially observed activities. 3) the similarity between partially observed actions is enlarged compared to complete activities.

To mitigate this issue, this chapter proposes a MSR framework, where multiple score functions are collaboratively learned corresponding to each performance stage. Partially observed activities in performance stages are collected at the learning stage to improve the robustness of score functions. Furthermore, the inherent evaluation of segments from adjacent performance stages is considered by introducing a soft label strategy into the learning formulation. By doing this, the MSR method is capable of identifying activities from partial observations with an outstanding performance.

6.2.2 Problem Statement

This chapter aims to develop an online activity recognition method for identifying ongoing activity sequences. The method is particularly designed to deal with the partial activity observation problem, and it is required to be robust to any activity segment.

Based on the evolution of an activity along the time domain, an activity is progressively divided into three stages: *start*, *peak*, and *end*. The *start* segments are in an onset stage describing the transition from the initial status to the *peak* status which includes the most salient information of an activity of interest, while the *end* segments are in an offset stage depicting the transition from the *peak* status back to the *end* status. The definition of performance stages can be application dependent. Given a fully observed activity sequence $X[1 : T]$ of length T, an arbitrary partial observation of it can be represented by $X[t_1 : t_2]$, where $1 \leq t_1 < t_2 \leq T$. Herein, t_1 is not constrained to be 1, which means that the partial activity could be at any performance stage. This overcomes the assumption that the partial activity needs to be observed from the start of an activity in (Escalante *et al.*, 2016; Ryoo, 2011). Note, T might vary in different activity sequences. The goal is to extract discriminative information for any partial activity $X[t_1 : t_2]$ and then assign an activity label to it.

6.2.3 3D Spatio-temporal Activity Representation

Recently, a lot of human activity recognition approaches based on skeleton joints have achieved satisfactory performance, owing to the invariance of the skeleton information to different locations and human appearances (Shotton *et al.*, 2013b). Here, the structured feature descriptor GBSW proposed in Chapter 3 is adapted for the online action recognition scenario. The coordinates of skeleton joints are firstly transformed with respect to a person-centric coordinate system. This transformation reduces the influence of various locations and orientations between subjects and the sensor. The histogram of moving trend (as shown in Fig. 6.3) is mined from the frame moving directions $\mathbf{V}_t$, which can effectively infer the intention of an activity at a high level. The directions in 3D space are empirically decomposed into l semantic moving words

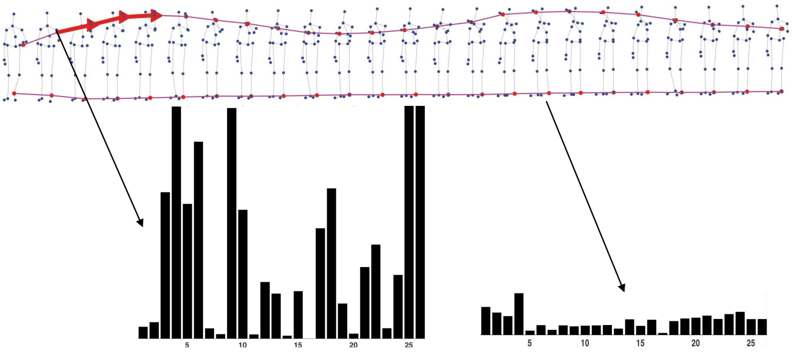


Fig. 6.3: The moving trend of skeleton joints.

$\overline{\mathbf{V}}_j, 1 \leq j \leq l$. Finally, the motion feature over a sequence is coded to the moving trend using the *cosine* similarity. For geometry feature, it is improved to make it adaptive to the segment-wise recognition by using the start of the segment as initial status x_{in}, y_{in}, z_{in}. Therefore, the process of feature extraction is summarized as follows:

$$\begin{cases} \mathbf{v_t^i} = \{x_{p_t^i} - x_{p_{t-1}^i}, y_{p_t^i} - y_{p_{t-1}^i}, z_{p_t^i} - z_{p_{t-1}^i}\} \\ cos\theta_j^i(t) = \frac{\mathbf{v_t^i} \cdot \overline{\mathbf{v}}_\mathbf{j}}{\|\mathbf{v_t^i}\| \|\overline{\mathbf{v}}_\mathbf{j}\|}, j \in [1, l] \\ bin_j = \sum_{t=t1}^{t2} \|\mathbf{v}_t^i\| \times max\{cos\theta_j^i(t)\}, j \in [1, l] \\ H(i) = \{bin_1, ..., bin_m\} \\ \triangle d_t^i = \{x_{p_t^i} - x_{in}^i, y_{p_t^i} - y_{in}^i, z_{p_t^i} - z_{in}^i\} \\ G(t) = \{\triangle d_t^1, ..., \triangle d_t^N\} \end{cases} \tag{6.1}$$

where $x_{p_t^i}, y_{p_t^i}, z_{p_t^i}$ represent the transformed coordinates of the $i-$th joint. $\mathbf{v_t^i}$ and $\overline{\mathbf{v}}_\mathbf{j}$ denote the frame-level moving direction and the semantic moving word. $H(i)$ and $G(t)$ mean the extracted moving trend feature and geometry feature, respectively.

6.2.4 Multi-stage Soft Regression

Let $(X_1,\boldsymbol{y}_1), (X_2,\boldsymbol{y}_2), ..., (X_n,\boldsymbol{y}_n)$ be the training data, where X_i is the $i-th$ activity sample and $\boldsymbol{y}_i$ is the label vector. To cope with random activity fragment $X[t_1 : t_2]$, the segment $x_i^{k,j} (x_i^{k,j} \in X_i)$ at each performance stage k from each category i is generated for learning, as shown in Fig. 6.2. A preliminary idea is to learn a single score function to measure the compatibility between activity fragments and labels using all segments, as denoted by the following formula:

$$\begin{gathered}\min_{\boldsymbol{W}_o} \sum_{i=1}^{n} \sum_{k=1}^{3} \sum_{j=1}^{m} ||\boldsymbol{W}_o^T \Phi(x_i^{k,j}) - \boldsymbol{y}_i||_{1,2} + \frac{\xi}{2} ||\boldsymbol{W}_o||_2^2 \\ s.t. \quad \xi \geq 0. \end{gathered} \tag{6.2}$$

where m is the number of activity segments from the performance stage k, $\boldsymbol{W}_o$ is the score function, and $\Phi(\cdot)$ is the proposed feature extraction function. This preliminary method is referred as Multi-stage Regression (MR) since activity segments at multiple stages are used for the regression function. The MR method has the capacity of identifying the partial activity when it happens at the *peak* stage as it includes the most salient information of an activity of interest. However, the power of discriminating similar segments from the *start* or *end* stage is insufficient, which will be discussed in Section 6.3.

In addition, an enhanced regression framework, named Multi-stage Soft Regression is introduced. This is a fine-grained regression framework, where multiple score functions for each specific performance stage are collaboratively learned to improve the power of discriminating similar activity segments. To remain the consistency among sequential segments, a soft label strategy between adjacent stages is implemented at learning time. As shown in Fig. 6.2, $\boldsymbol{W}_s^T, \boldsymbol{W}_p^T, \boldsymbol{W}_e^T$ are score functions corresponding to each performance stage, which are collaboratively learned by soft labelling segments from adjacent performance stages. The color gradient indicates the general trend of the information of an activity along the time domain. Segments in the dark color contain more information than segments in the light color, and the information becomes more while the color becoming darker. The MSR method is formulated as follows:

$$\begin{aligned}
&\min_{\boldsymbol{W},\lambda} \sum_{i=1}^{n}\sum_{k=1}^{2}\sum_{j=1}^{m} ||\boldsymbol{W}_s^T \Phi(x_i^{k,j}) - \boldsymbol{y}_i(1-\lambda_{s,j}|Sgn(k-1)|)||_{1,2} \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{3}\sum_{j=1}^{m} ||\boldsymbol{W}_p^T \Phi(x_i^{k,j}) - \boldsymbol{y}_i(1-\lambda_{p,j}|Sgn(k-2)|)||_{1,2} \\
&+ \sum_{i=1}^{n}\sum_{k=2}^{3}\sum_{j=1}^{m} ||\boldsymbol{W}_e^T \Phi(x_i^{k,j}) - \boldsymbol{y}_i(1-\lambda_{e,j}|Sgn(k-3)|)||_{1,2} \\
&+ \frac{\xi_s}{2}||\boldsymbol{W}_s||_2^2 + \frac{\xi_p}{2}||\boldsymbol{W}_p||_2^2 + \frac{\xi_e}{2}||\boldsymbol{W}_e||_2^2 \\
&s.t. \quad \xi_s, \xi_p, \xi_e \geq 0, \quad 0 \leq \lambda_{s,j}, \lambda_{p,j}, \lambda_{e,j} \leq 1
\end{aligned} \tag{6.3}$$

where $\boldsymbol{W}_s^T$, $\boldsymbol{W}_p^T$, $\boldsymbol{W}_e^T$ are score functions corresponding to the *start*, *peak*, and *end* stage, respectively. $Sgn(\cdot)$ is the sign function and m is the number of segments in each stage. $(1-\lambda_{s,j}|Sgn(k-1)|)$ is the weight of segments to ensure stronger labels for segments from k stage than that of its adjacent stage. For simplification, Eq.6.3 is rewritten as follows:

$$\begin{aligned}
&\min_{\boldsymbol{W},\theta} \sum_{i=1}^{n} ||\boldsymbol{W}_s^T \widehat{\mathbf{X}}_i - \boldsymbol{Y}_i\theta_{sp}||_{1,2} + \frac{\xi_s}{2}||\boldsymbol{W}_s||_2^2 \\
&+ ||\boldsymbol{W}_p^T \widehat{\mathbf{X}}_i - \boldsymbol{Y}_i\theta_{spe}||_{1,2} + \frac{\xi_p}{2}||\boldsymbol{W}_p||_2^2 \\
&+ ||\boldsymbol{W}_e^T \widehat{\mathbf{X}}_i - \boldsymbol{Y}_i\theta_{pe}||_{1,2} + \frac{\xi_e}{2}||\boldsymbol{W}_e||_2^2 \\
&s.t. \quad \xi_s, \xi_p, \xi_e \geq 0, \quad 0 \leq \theta_{sp}, \theta_{spe}, \theta_{pe} \leq 1
\end{aligned} \tag{6.4}$$

where $\widehat{\mathbf{X}}$ denotes the feature descriptor matrix of segments, and $\boldsymbol{Y}$ is the corresponding label matrix. The elements in θ_{sp} corresponding to the segments at the *start* stage are constrained to be 1 while learning $\boldsymbol{W}_s^T$ (by analogy, θ_{spe} and θ_{pe} have the same affinity).

Compared to MR, there are three score functions $\boldsymbol{W}_s^T$, $\boldsymbol{W}_p^T$, $\boldsymbol{W}_e^T$ collaboratively learned in this MSR regression framework.

6.2.5 Optimization

The optimization problem Eq. 6.4 is solved by iteratively optimizing specific parameters at each step while holding the others fixed. The details are shown below:

Step 1. Fix θ and optimize $\boldsymbol{W}$: the gradient of Eq. 6.4 with respect to $\boldsymbol{W}$ can be represented by:

$$Gradient_w = \sum_{i=1}^{n} \frac{\partial ||\boldsymbol{W}^T\widehat{\mathbf{X}}_i - \boldsymbol{Y}_i\theta||_{1,2}}{\partial \boldsymbol{W}} + \xi_s \boldsymbol{W}, \tag{6.5}$$

then the updated parameter at each time step is given by:

$$\boldsymbol{W}(t) = \boldsymbol{W}(t-1) - \tau Gradient_w, \tag{6.6}$$

where τ is the iteration step size.

Step 2. Fix $\boldsymbol{W}$ and optimize θ: θ is solved by the standard gradient descent method. Specifically, the gradient of Eq. 6.4 with respect to θ is given by:

$$Gradient_\theta = \sum_{i=1}^{n} \frac{\partial ||\boldsymbol{W}^T\widehat{\mathbf{X}}_i - \boldsymbol{Y}_i\theta||_{1,2}}{\partial \theta}, \tag{6.7}$$

then the updated at each time step is given by:

$$\theta(t) = \theta(t-1) - \mu Gradient_\theta \tag{6.8}$$

is projected into the constrained space. Here μ is the iteration step size.

Please note that $\boldsymbol{W}$ could be $\boldsymbol{W}_s$, $\boldsymbol{W}_p$, and $\boldsymbol{W}_e$, and θ could be θ_{sp}, θ_{spe}, and θ_{pe}.

6.2.6 Activity Fusion

In a practical video stream, it is difficult to decide the exact performance stage of a partially observed activity because of the ambiguous boundary among observations. Therefore, instead of using a single score function to identify it, the results from three score functions are fused using a Gaussian function.

$$\underset{label_n}{\arg\max} \sum_{n=1}^{c} \sum_{k=s,p,e} G_k \boldsymbol{W}_k^T(label_n,:)\Phi(\mathbf{x}_i) \tag{6.9}$$

here,

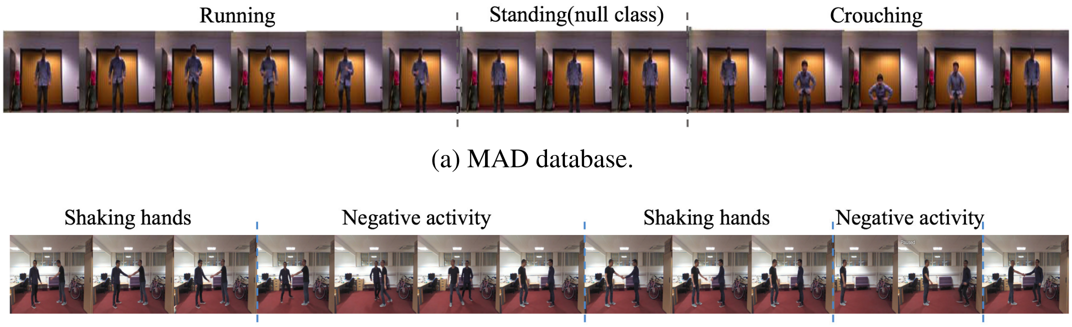
$$G_k = \frac{1}{\sqrt{2\pi}} exp(-\frac{(\Phi(\mathbf{x}_i) - \mu_k)^2}{2\delta_k^2}) \tag{6.10}$$

where $\boldsymbol{W}_k^T(label_n,:)$ is the learned coefficients of category $label_n$, c is the number of activity classes, and G_k is the weight of the score produced by $\boldsymbol{W}_k^T(label_n,:)$. G_k is calculated by a Gaussian function with mean μ_k and standard deviation δ_k of observations at the performance stage $k(k \in (s,p,e))$.

6.3 Experimental Evaluation

6.3.1 Introduction of Datasets

The proposed method is evaluated on the public MAD dataset (Huang *et al.*, 2014) and the newly collected OHI dataset. The MAD database has 40 sequences performed by 20 subjects (2 sequences each subject). Each sequence contains 35 actions continuously performed by one subject. The OHI dataset contains total 10 human interactions and each interaction is performed by 23 pairs of subjects. Fig. 6.4a and Fig. 6.4b show some sample frames and two datasets. In the MAD dataset, the time series between two actions are considered as the null class where the subject keeps standing in most cases, while in the OHI dataset, the time series between two activities (subjects are performing as they want) is considered as the neutral activities, which makes the OHI dataset more challenging due to the larger variability of the neutral activities.



(a) MAD database.

(b) OHI database.

Fig. 6.4: Sample frames of the *MAD database* and *OHI database*.

6.3.2 Evaluation of MSR Performance

The preliminary method MR and its improved version MSR are proposed to handle the partial activity observation problem in online activity recognition. Compared to the single score function in MR, multiple score functions are collaboratively learned via a soft label strategy in the MSR framework. This experiment intends to evaluate the improvement of MSR in terms of the property of discriminating similar partial activities and the robustness to arbitrary activity segments. Complete activities in both MAD and OHI databases are divided into three performance stages, from which overlapped activity segments are selected.

Table 6.1 shows the recognition performance of the proposed methods on two databases. From the first row in the table, it can be found that the recognition accuracies on the *start* and the *end* segments are much lower than that of the *peak* segments in MR. This investigation verifies a fact that segments from the *peak* period are easier to be identified because their information is more discriminative than both *start* and *end* stages. Observing that the necessity of improving the performance of partial activity recognition, especially for segments from the *start* and *end* stages, the soft label between adjacent stages is introduced in MSR to consider the inherent evolution of activities. The comparison between MR and MSR is given to demonstrate the enhanced recognition ability of MSR on similar segments.

Table 6.1: Recognition Accuracy (%) of partial activities from the *MAD database* and *OHI database*.

	MAD database				OHI database			
	start	*peak*	*end*	**Average**	*start*	*peak*	*end*	**Average**
MR	67.87	70.75	55.35	**64.66**	42.51	61.40	38.46	**47.46**
MSR	80.28	83.33	65.88	**76.50**	54.74	76.53	60.00	**63.76**

The second row in Table 6.1 reports that the MSR method significantly improves the overall performance by increasing the average accuracies from 64.66% to 76.50% and from 47.46% to 63.76% on MAD database and OHI database, respectively. This achievement owes much to the collaboratively learned score functions with a focus on specific performance stages, which strengthens the recognition power of functions $\boldsymbol{W}_s^T$ and $\boldsymbol{W}_e^T$. Specifically, for two databases, the recognition accuracies of MR on three stages are improved over 10% by MSR. More obviously, MSR achieves 60% on the *end* stage on OHI database, which is approximately 20% higher than that of MR.

Fig. 6.5 shows the confusion matrices of the MR and MSR methods on the OHI database. In the MR method, the recognition accuracies of observations from the *start* and *end* stage in most activity classes are less than 50%, and even for the observations from the *peak* stage which posses relatively rich characteristics of each activity category. For example, activities such as *punching*, and *hugging* only have 47.1% and

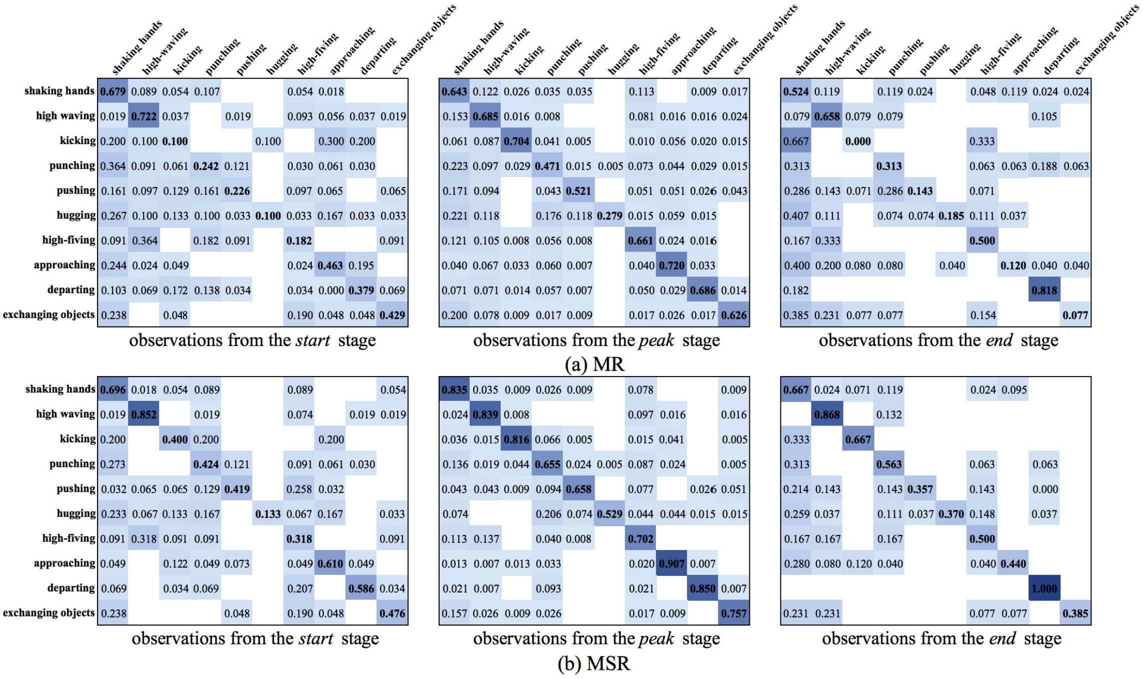


Fig. 6.5: Confusion matrixes of MR and MSR for observations from the *start, peak, end* performance stage on the OHI database.

27.9% accuracy mainly due to the confusion with *shaking hands*. This poor performance, especially for the beginning and ending parts, is mainly caused by the limited information from minor temporal activity sequences. The MSR method significantly boosts the overall recognition performance by learning a specific score function for each performance stage as well as considering the inherent evolution of activities via the soft label strategy. For example, MSR can correctly classify over 80% of the observations from the *peak* stage, and it dramatically improves the accuracy of *kicking* by 66.7%.

6.3.3 Comparative Results in Online Activity Recognition

Since the target of online activity recognition is to detect and recognize activities as soon as possible in a continuous video stream, this experiment therefore tests MSR's capacity of identifying activities from continuous video streams. The online operation is executed using a sliding window which sequentially selects the activity segments over the time. Each segment will produce its own preliminary activity class by fusing the results from the multiple score functions via the proposed activity fusion method. Then a consistency among the sequential segments will be yielded by a segmental smoothing function.

For the MAD database, the proposed MSR method is compared with the existing state-of-the-art methods in terms of *Prec* and *Rec*. The same experimental settings and definitions given in (Huang *et al.*, 2014) are used: *TN*: the number of correctly detected activities who have 50% overlap with the ground truth activity; *DN*: the number of detected activities; *GTN*: the number of all ground truth activity classes, $Prec = \frac{TN}{DN}$ and $Rec = \frac{TN}{GTN}$. From Table 6.2, it is noticed that MSR has the best performance of over 80% accuracy in both *Prec* and *Rec* among all methods, which indicates its capacity of precisely detect activity from videos and accurately classifying the detected activities.

Table 6.2: Online Performance: Average Detection and Recognition results on the *MAD database* (%)

Methods	*Prec*	*Rec*
SVM+DP(Hoai *et al.*, 2011)	28.6	51.4
SMMED(Huang *et al.*, 2014)	59.2	57.4
ENB(Escalante *et al.*, 2016)	76.1	73.6
(Devanne *et al.*, 2017b)	72.1	79.7
Beyond Joints+RNN(Wang & Wang, 2018)	74.2	73.4
MSR	**81.3**	**82.3**

Fig. 6.6 shows the online performance on two test sequences. For each test sequence, the ground truth labels, results of SVM+DP (Hoai *et al.*, 2011), SMMED (Huang *et al.*, 2014), ENB (Escalante *et al.*, 2016), and MSR are listed. It can be observed that although SVM+DP can detect the occurrence of each activity in two test sequences, it has frequently fragmented labels around ground truth labels, meaning the low recognition rate on detected activities. Though SMMED and ENB improve the recognition performance, they suffer from the increasing rates of missing detection. The colored bars of the MSR method in two sequences are the closest to the ground truth bars by no fragmented labels and higher classification accuracies.

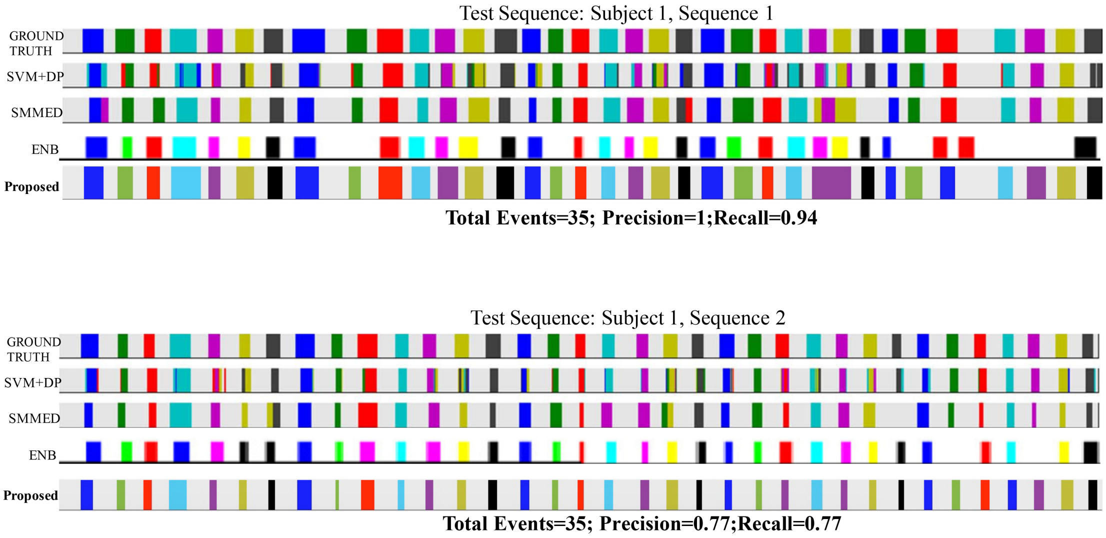


Fig. 6.6: The comparison of MSR with SVM+DP, SMMED, and ENB on two test sequences from the MAD database.

In addition, to evaluate the proposed MR and MSR methods in terms of the robustness to arbitrary activity fragments in the online scenario, the five-fold-cross-validation is performed on the MAD database. The 20 subjects in MAD database are divided into 5 groups and each group contains 4 subjects. The average accuracies are obtained by successively using the sequences from 4 groups as training and the sequences from the remaining group as testing. At testing time, the activity segments are possible at any performance stage. Fig. 6.7 reports the average recognition accuracy of each activity category. The figure shows that MSR outperforms MR on most of the activities. Especially, MSR achieves statistically significant improvements on *Both Arms Pointing to Right Side* and *Right Arm Punch*, by increasing accuracies from 13.13% to 100% and from 17.5% to 94.72%, respectively. Besides, compared to MR, MSR has smaller standard deviations of results in all activity categories. This demonstrates that the adaptability to random activity segments is strengthened by the collaboratively learned score functions in MSR.

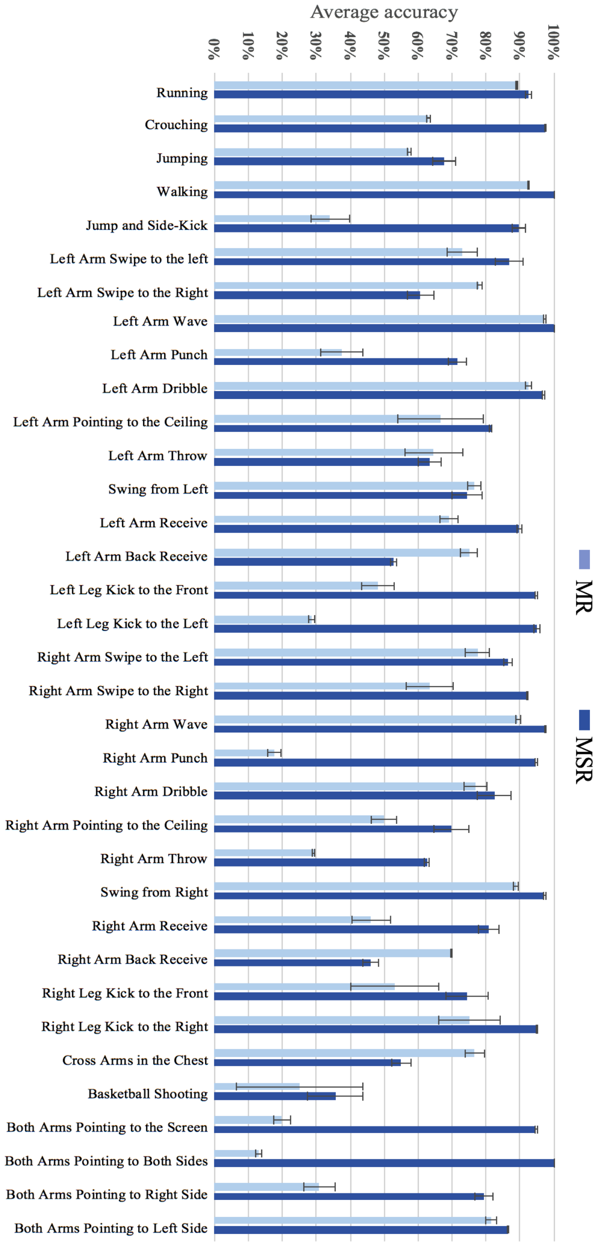


Fig. 6.7: Average recognition accuracy on each activity category. Error bars indicate standard deviations.

6.4 Summary

This chapter proposed a multi-stage soft regression framework to deal with the partial activity observation problem in online activity recognition. The MSR framework delicately assembled overlapped activity observations in any period and also considered the relation between adjacent performance stages to improve its robustness to arbitrary activity segments. By formulating the online activity recognition task as a multi-stage soft regression problem, multiple score functions were collaboratively learned to effectively discriminate similar partial activity fragments.

Compared to MR, the improvements of MSR were firstly validated by over 10% increases of the accuracies in each performance stage on both MAD and OHI datasets, owing to reducing the confusion among activities. Furthermore, the MSR method achieved an accuracy of 82.3% on the MAD database with a significant improvement by 10.2% over the state-of-the-art method.

Although online activity recognition is essential in practical applications, it is still a problem far from being solved, as artificial settings in both algorithms and databases are far away from realistic scenarios, for example, all videos in the OHI dataset are captured from a specific viewpoint. Therefore, extending the database with multiple viewpoints and improving the proposed method to make it invariant to different viewpoints could be the future work.

Chapter 7

Conclusions

This chapter provides a summary of the contributions of this thesis and identifies the future work of human activity analysis.

7.1 Overview

The main objective of this thesis is to develop effective human activity recognition algorithms, which can be applied in many aspects such as human-robot interaction, healthcare, video surveillance, elderly care, and education.

The emergency of RGB-D sensors provide an easy access to 3D information of human in scenes by providing the extra depth information and skeleton joints. This technology has greatly boosted the development of human activity recognition using RGB-D data. To this end, this thesis has reviewed various human action and interaction recognition methods, which were divided into three categories according to data modality. Many approaches have been proposed to understand actions performed by a single person. However, human interactions which are more complex than human actions due to more variations in subject appearance, scale, viewpoint, interacting motion patterns are less explored. Though some methods achieved high recognition rate, most of them are offline and their online performance which needs to simultaneously detect activities from continuous videos and identify the activity type remains unclear.

To address the above problems, this thesis firstly proposed an effective spatio-temporal feature descriptor GBSW which aggregates the BSW with the G feature to describe human actions from skeleton sequences. The proposed BSW and G could extract the 3D moving trend and motion cues of skeleton sequences in the spatial and temporal domains, respectively. Secondly, moving similarity between body parts was extracted from interactive persons to describe their mutual relationship for human interaction recognition. Also, the new OHI dataset was collected to be served as a benchmark for the future research. Thirdly, a skeleton motion distribution based approach was developed to detect activities in continuous videos based on the change of the motion distribution from each other. Finally, to deal with the partial activity observation problem in online activity recognition, the MSR framework was introduced to collaboratively learn multiple score functions with a focus on each performance stage to improve the recognition accuracy.

7.2 Contributions

This section presents the main contributions of this thesis which include the GBSW method for accurate human action recognition, the SRMS approach for human interaction, the SMD model for activity detection, and the MSR framework for effective online activity recognition.

In Chapter 3, an effective spatio-temporal feature descriptor GBSW was proposed for human action recognition using skeleton sequences. It aggregated the BSW with the G feature to describe human actions from skeleton sequences. The proposed BSW was constructed to describe the moving trend of skeleton joints in 3D space. Meanwhile, the G feature along the whole action sequence was calculated to extract motion cues in the temporal domain. Outstanding performance over the state-of-the-art methods on public datasets was achieved by the GBSW, owing to the semantic representation and the complementary effect of the aggregation of different types of features.

In Chapter 4, the moving similarity between body parts was further mined by taking advantage of the BSW feature extracted from individuals for human interaction recognition. In addition, an RGB-D based OHI dataset was collected to serve as the benchmark for the evaluation of human interaction recognition algorithms. The experiment results on both SBU Kinect Interaction dataset and OHI dataset have proven the effectiveness of the proposed method in discriminating complex human interactions.

In Chapter 5, the SMD model was proposed for action detection in continuous videos. The occurrence of actions was detected depending on the change of the motion distribution in different action categories. Once an action was detected, a snippet-based classifier was designed to process the observed video immediately for action classification. Experimental results on public datasets have demonstrated that the proposed method can effectively detect human actions in continuous videos.

In Chapter 6, the MSR framework was constructed to address the partial activity observation problem in online activity recognition. Multiple score functions with a focus on each performance stage were collaboratively learned to make it robust to arbitrary activity fragments. The inherent evolution of partial activities from adjacent stages was also modeled by introducing a soft label strategy into the learning formulation. Extensive experimental results on the MAD database and the OHI database have demonstrated the superior performance of the MSR method over the state-of-the-art approaches.

7.3 Future Work

This section discusses the limitations of our work and some directions for the future work.

7.3.1 Fusion of Multi-modal Data

The effectiveness and efficiency of the proposed GBSW method and the SRMS method has been demonstrated through experiments on recognizing human actions and interactions from human skeleton sequences. These methods are compact and straightforward to depict the motion properties of actions. However, they lack the capacity of describing the appearance or the contextual information from the surrounding environment which is important for understanding complex activities (e.g., human-human, human-object, and human-environment interactions). Thus, combining the proposed methods with extra features from other modalities, such as depth and RGB which have the outstanding capability to describe the appearance cues, will be a future research direction.

7.3.2 Improvement of Adaptiveness to Different Tasks

Developing an action recognition method for a specific scene or action database is relative simple. However, it remains a great challenge to develop a universal action recognition methods for different tasks since the requirements of action types and recognition performance various to a large extend in different situations. Also, it is labor-intensive to repeatedly collect large amount of action samples and develop specific models for each application. Thus, improving the adaptive feature and the robustness of the proposed MSR framework to different real-world environments will be another future research direction.

7.3.3 Improvement of Effectiveness in Activity Detection

With the quick prevalence of vision sensors on both public and private areas, a large amount of human activity related video data has been collected nowadays. The performance of existing activity recognition methods is very likely to be improved by training on these large data. However, most of the existing methods require manually segmentation and labeling operations on the recorded videos for training purpose. To solve this problem, It is attracting to employ existing activity detection methods to automatically labeling a large amount of video data. Thus, improve the effectiveness of activity detection is a promising future direction.

7.3.4 Improvement of Performance in Online Activity Recognition

Online activity recognition aims to instantly detect and recognize human actions from continuously input videos. It is a challenging task, especially in unconstrained scenes, where a large number of negative activities make the detection and recognition of expected actions difficult. The outstanding performance of the developed MSR framework has been evaluated through both detection and recognition accuracy measurements. However, it still remains a great challenge to recognize an ongoing action correctly in an instant performance. Thus, improving the efficiency of the proposed MSR framework to the real-world environment while at the same time keeping good recognition accuracy will be another future research direction.

7.3.5 Towards Deep Understanding of Human Behavior

Human behavior identification requires a long period observation of human activities and might evolve some other information such as human affective status and visual focus of attention. Many practical applications require a deep understanding of human behavior in order to make a robust judgement. For example, in a public safety monitoring system, a criminal activity might have some prior wondering cues or several specificity types of attention models. The safety system should avoid generating a warning for similar activities such as running or falling. In a therapeutic intervention with children with ASD, the labor of therapists will be reduced by the automatic analysis of children's behavior. The long-term analyzed data is also beneficial in producing a consistent therapeutic intervention experience. Thus, it is a promising future direction to develop methods for a deep understanding of human behavior.

References

AGGARWAL, J.K. & RYOO, M.S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, **43**, 16. 1

ALAZRAI, R., MOWAFI, Y. & LEE, C.G. (2015). Anatomical-plane-based representation for human–human interactions analysis. *Pattern Recog.*, **48**, 2346–2363. 21

ALTHLOOTHI, S., MAHOOR, M.H., ZHANG, X. & VOYLES, R.M. (2014). Human activity recognition using multi-features and multiple kernel learning. *Pattern Recog.*, **47**, 1800–1812. 16, 37

ARGYRIOU, V., PETROU, M. & BARSKY, S. (2010). Photometric stereo with an arbitrary number of illuminants. *Computer Vision and Image Understanding*, **114**, 887–900. 8

BARADEL, F., WOLF, C. & MILLE, J. (2017). Pose-conditioned spatio-temporal attention for human action recognition. *arXiv preprint arXiv:1703.10106, 2017*. 66, 67

BARLA, A., FRANCESCHI, E., ODONE, F. & VERRI, A. (2002). Image kernels. *Pattern Recognition with Support Vector Machines*, 83–96. 59

BENGALUR, M.D. (2013). Human activity recognition using body pose features and support vector machine. In *Int. Conf. Advances in Computing, Commun. Inform.*, 1970–1975. 21

BLOOM, V., ARGYRIOU, V. & MAKRIS, D. (2013). Dynamic feature selection for online action recognition. In *Int. Workshop on Human Behavior Understanding*, 64–76. 30

BLOOM, V., ARGYRIOU, V. & MAKRIS, D. (2014). G3di: A gaming interaction dataset with a real time detection and evaluation framework. In *Proc. Eur. Conf. Comput. Vis. Workshop*, 698–712. 20, 34, 64

BLOOM, V., ARGYRIOU, V. & MAKRIS, D. (2016). Hierarchical transfer learning for online recognition of compound actions. *Comput. Vis. Image Understanding*, **144**, 62–72. 20

BLOOM, V., ARGYRIOU, V. & MAKRIS, D. (2017). Linear latent low dimensional space for online early action recognition and prediction. *Pattern Recognition*, **72**, 532–547. 30

BLUNSDEN, S. & FISHER, R. (2010). The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, **4**, 4. 33

BOIMAN, O., SHECHTMAN, E. & IRANI, M. (2008). In defense of nearest-neighbor based image classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1–8. 16

BULBUL, M.F., JIANG, Y. & MA, J. (2015). Human action recognition based on dmms, hogs and contourlet transform. In *Proc. IEEE Int. Conf. Multimedia Big Data*, 389–394. 14, 15

CAI, X., ZHOU, W., WU, L., LUO, J. & LI, H. (2016). Effective active skeleton representation for low latency human action recognition. *IEEE Trans. Multimedia*, **18**, 141–154. 37, 81, 83

CHAARAOUI, A.A., PADILLA-LÓPEZ, J.R., CLIMENT-PÉREZ, P. & FLÓREZ-REVUELTA, F. (2014). Evolutionary joint selection to improve human action recognition with rgb-d devices. *Expert systems with applications Expert Syst. Appl.*, **41**, 786–794. 13

CHAI, X., LIU, Z., YIN, F., LIU, Z. & CHEN, X. (2016). Two streams recurrent neural networks for large-scale continuous gesture recognition. In *Int. Conf. Pattern Recog. Workshops*, 31–36. 31

CHANG, C.C. & LIN, C.J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**, 27. 50, 66

CHEN, C., LIU, K. & KEHTARNAVAZ, N. (2013). Real-time human action recognition based on depth motion maps. *J. Real-Time Image Processing*, 1–9. 14

CHEN, C., LIU, M., LIU, H., ZHANG, B., HAN, J. & KEHTARNAVAZ, N. (2017). Multi-temporal depth motion maps-based local binary patterns for 3-d human action recognition. *IEEE Access*, **5**, 22590–22604. 37

COPPOLA, C., FARIA, D.R., NUNES, U. & BELLOTTO, N. (2016). Social activity recognition based on probabilistic merging of skeleton features with proximity priors from rgb-d data. In *IEEE/RSJ Int. Conf. Intell. Robots and Systems*, 5055–5061. 21, 64

CRABBE, B., PAIEMENT, A., HANNUNA, S. & MIRMEHDI, M. (2015). Skeleton-free body pose estimation from depth images for movement analysis. In *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 70–78. 27

DE GEEST, R., GAVVES, E., GHODRATI, A., LI, Z., SNOEK, C. & TUYTELAARS, T. (2016). Online action detection. In *European Conference on Computer Vision*, 269–284, Springer. 70

DENG, L., LEUNG, H., GU, N. & YANG, Y. (2012). Generalized model-based human motion recognition with body partition index maps. In *Comput. Graphics Forum*, vol. 31, 202–215. 12

DEVANNE, M., WANNOUS, H., BERRETTI, S., PALA, P., DAOUDI, M. & DEL BIMBO, A. (2015). 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Trans. Cybern.*, **45**, 1340–1352. 37, 53, 54, 55

DEVANNE, M., BERRETTI, S., PALA, P., WANNOUS, H., DAOUDI, M. & DEL BIMBO, A. (2017a). Motion segment decomposition of rgb-d sequences for human behavior understanding. *Pattern Recog.*, **61**, 222–233. 29, 78

DEVANNE, M., BERRETTI, S., PALA, P., WANNOUS, H., DAOUDI, M. & DEL BIMBO, A. (2017b). Motion segment decomposition of rgb-d sequences for human behavior understanding. *Pattern Recog.*, **61**, 222–233. 92

DU, Y., FU, Y. & WANG, L. (2015a). Skeleton based action recognition with convolutional neural network. In *IAPR Asian Conf. Pattern Recog.*, 579–583. 24

DU, Y., WANG, W. & WANG, L. (2015b). Hierarchical recurrent neural network for skeleton based action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1110–1118. 25, 37, 67, 70

DU, Y., FU, Y. & WANG, L. (2016). Representation learning of temporal dynamics for skeleton-based action recognition. *IEEE Trans. Image Process.*, **25**, 3010–3022. 25

ELGAMMAL, A., DURAISWAMI, R., HARWOOD, D. & DAVIS, L.S. (2002). Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, **90**, 1151–1163. 71

ESCALANTE, H.J., MORALES, E.F. & SUCAR, L.E. (2016). A naive bayes baseline for early gesture recognition. *Pattern Recognition Letters*, **73**, 91–99. 78, 84, 92

EVANGELIDIS, G.D., SINGH, G. & HORAUD, R. (2014). Continuous gesture recognition from articulated poses. In *Proc. Eur. Conf. Comput. Vis. Workshops*, 595–607. 81

EWEIWI, A., CHEEMA, M.S., BAUCKHAGE, C. & GALL, J. (2015). Efficient pose-based action recognition. In *Asian Conf. Comput. Vis.*, 428–443. 13

FOTHERGILL, S., MENTIS, H., KOHLI, P. & NOWOZIN, S. (2012). Instructing people for training gestural interactive systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1737–1746. 30

GAGLIO, S., RE, G.L. & MORANA, M. (2015). Human activity recognition process using 3-d posture data. *IEEE Trans. Human-Mach. Syst.*, **45**, 586–597. 40

GONG, D., MEDIONI, G. & ZHAO, X. (2014). Structured time series analysis for human action segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell*, **36**, 1414–1427. 30

GORI, I., AGGARWAL, J. & RYOO, M.S. (2015). Building unified human descriptors for multi-type activity recognition. *CoRR, abs/1507.02558, 2015*, **3**. 67

GORI, I., AGGARWAL, J., MATTHIES, L. & RYOO, M.S. (2017). Multi-type activity recognition from a robot's viewpoint. In *Proc. Int. Joint Conf. Artificial Intell.*, 4849–4853. 19, 21

GOWAYYED, M.A., TORKI, M., HUSSEIN, M.E. & EL-SABAN, M. (2013). Histogram of oriented displacements (hod): describing trajectories of human joints for action recognition. In *Proc. Int. joint Conf. Artificial Intell.*, 1351–1357. 12, 40, 54

GRAVES, A., FERNÁNDEZ, S., GOMEZ, F. & SCHMIDHUBER, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. Int. Conf. Mach. Learn.*, 369–376. 30

GUO, Y., LI, Y. & SHAO, Z. (2017). Dsrf: A flexible trajectory descriptor for articulated human action recognition. *Pattern Recog.*. 70, 81

GUO, Y., LI, Y. & SHAO, Z. (2018). Dsrf: A flexible trajectory descriptor for articulated human action recognition. *Pattern Recog.*, **76**, 137–148. 13, 37

HAN, F., REILY, B., HOFF, W. & ZHANG, H. (2017). Space-time representation of people based on 3d skeletal data: A review. *Comput. Vis. Image Understanding*, **158**, 85–105. 11

HE, X. & NIYOGI, P. (2003). Locality preserving projections. **16**. 72

HOAI, M. & DE LA TORRE, F. (2014). Max-margin early event detectors. *International Journal of Computer Vision*, **107**, 191–202. 70, 83

HOAI, M., LAN, Z.Z. & DE LA TORRE, F. (2011). Joint segmentation and classification of human actions in video. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 3265–3272. 78, 92

HOCHREITER, S. & URGEN SCHMIDHUBER, J. (1997). Long Short-Term Memory. *Neural Comput.*, **9**, 1735–1780. 25

HOU, Y., LI, Z., WANG, P. & LI, W. (2018). Skeleton optical spectra based action recognition using convolutional neural networks. *IEEE Trans. Circuits Syst. Video Technol.*, **28**, 807–811. 24

HU, T., ZHU, X., GUO, W. & SU, K. (2013). Efficient interaction recognition through positive action representation. *Math. Problems in Eng.*, **2013**. xi, 20, 21, 64

HUANG, D., YAO, S., WANG, Y. & DE LA TORRE, F. (2014). Sequential max-margin event detectors. In *Proc. Eur. Conf. Comput. Vis.*, 410–424. 5, 6, 29, 34, 64, 78, 83, 89, 92

HUANG, Z., WAN, C., PROBST, T. & VAN GOOL, L. (2017). Deep learning on lie groups for skeleton-based action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 6099–6108. 24

HUYNH-THE, T., BANOS, O., LE, B.V., BUI, D.M., LEE, S., YOON, Y. & LE-TIEN, T. (2015). Pam-based flexible generative topic model for 3d interactive activity recognition. In *IEEE Int. Conf. Advanced Technol. Commun.*, 117–122. 19

IJJINA, E.P. & CHALAVADI, K.M. (2017). Human action recognition in rgb-d videos using motion sequence information and deep learning. *Pattern Recog.*, **72**, 504–516. 27

JALAL, A., KIM, Y.H., KIM, Y.J., KAMAL, S. & KIM, D. (2017). Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recog.*, **61**, 295–308. 16, 37, 52, 54, 70, 81

JI, S., XU, W., YANG, M. & YU, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell*, **35**, 221–231. 28

JI, X., CHENG, J., FENG, W. & TAO, D. (2018). Skeleton embedded motion body partition for human action recognition using depth sequences. *Signal Processing*, **143**, 56–68. 16, 37

JI, Y., YE, G. & CHENG, H. (2014). Interactive body part contrast mining for human interaction recognition. *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, 1–6. 20, 38, 56, 67

JI, Y., CHENG, H., ZHENG, Y. & LI, H. (2015). Learning contrastive feature distribution model for interaction recognition. *J. Vis. Commun. Image Represent.*, **33**, 340–349. 20, 38, 56, 67

JIA, C. & FU, Y. (2016). Low-rank tensor subspace learning for rgb-d action recognition. 37

JONES, S. & SHAO, L. (2014). Unsupervised spectral dual assignment clustering of human actions in context. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 604–611. 22

KAMEL, A., SHENG, B., YANG, P., LI, P., SHEN, R. & FENG, D.D. (2018). Deep convolutional neural networks for human action recognition using depth maps and postures. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 1–14. 37

KE, Q., BENNAMOUN, M., AN, S., SOHEL, F. & BOUSSAID, F. (2017). A new representation of skeleton sequences for 3d action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 4570–4579. 23, 38

KONG, Y. & FU, Y. (2014). Modeling supporting regions for close human interaction recognition. In *Proc. Eur. Conf. Comput. Vis. Workshops*, 29–44. 3

KONG, Y. & FU, Y. (2015a). Bilinear heterogeneous information machine for rgb-d action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1054–1062. 17

KONG, Y. & FU, Y. (2015b). Bilinear heterogeneous information machine for rgb-d action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1054–1062. 70

KONG, Y. & FU, Y. (2016). Close human interaction recognition using patch-aware models. *IEEE Trans. Image Process.*, **25**, 167–178. 18

KONG, Y., JIA, Y. & FU, Y. (2012). Learning human interaction by interactive phrases. In *Proc. Eur. Conf. Comput. Vis.*, 300–313. 18

KONG, Y., KIT, D. & FU, Y. (2014). A discriminative model with multiple temporal scales for action prediction. 596–611. 70

KONG, Y., SATARBOROUJENI, B. & FU, Y. (2016). Learning hierarchical 3d kernel descriptors for rgb-d action recognition. *Comput. Vis. Image Understanding*, **144**, 14–23. 37

KONIUSZ, P., CHERIAN, A. & PORIKLI, F. (2016). Tensor representations via kernel linearization for action recognition from 3d skeletons. In *Proc. Eur. Conf. Comput. Vis.*, 37–53. 37

KRIZHEVSKY, A., SUTSKEVER, I. & HINTON, G.E. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105. 7, 23

KULKARNI, K., EVANGELIDIS, G., CECH, J. & HORAUD, R. (2015). Continuous action recognition based on sequence alignment. *Int. J. Comput. Vis.*, **112**, 90–114. 29, 81

LATECKI, L.J., LAZAREVIC, A. & POKRAJAC, D. (2007). Outlier detection with kernel density functions. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, 61–75, Springer. 71, 72

LI, C., HOU, Y., WANG, P. & LI, W. (2017). Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Processing Lett.*, **24**, 624–628. 23

LI, W., ZHANG, Z. & LIU, Z. (2010). Action recognition based on a bag of 3d points. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 9–14. xi, 6, 14, 31, 32, 37, 49, 50, 52, 53, 54, 57

LI, W., WEN, L., CHOO CHUAH, M. & LYU, S. (2015). Category-blind human action recognition: a practical recognition system. In *Proc. IEEE Int. Conf. Comput. Vision*, 4444–4452. 67

Li, Y., Lan, C., Xing, J., Zeng, W., Yuan, C. & Liu, J. (2016). Online human action detection using joint classification-regression recurrent neural networks. In *Proc. Eur. Conf. Comput. Vis.*, 203–220. 31

Lillo, I., Niebles, J.C. & Soto, A. (2017). Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos. *Image Vis. Comput.*, **59**, 63–75. 13, 40, 54

Liu, A.A., Nie, W.Z., Su, Y.T., Ma, L., Hao, T. & Yang, Z.X. (2015). Coupled hidden conditional random fields for rgb-d human action recognition. *Signal Processing*, **112**, 74–82. 17

Liu, B., Yu, H., Zhou, X., Tang, D. & Liu, H. (2016a). Combining 3d joints moving trend and geometry property for human action recognition. In *IEEE Int. Conf. Syst. Man, Cyber.*, 000332–000337. 37, 52, 54, 55

Liu, B., Cai, H., Ji, X. & Liu, H. (2017a). Human-human interaction recognition based on spatial and motion trend feature. In *IEEE Int. Conf. Image Processing*, 4547–4551. 4, 38

Liu, C., Hu, Y., Li, Y., Song, S. & Liu, J. (2017b). Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475, 2017*. 35

Liu, J., Shahroudy, A., Xu, D. & Wang, G. (2016b). Spatio-temporal lstm with trust gates for 3d human action recognition. In *Proc. Eur. Conf. Comput. Vis.*, 816–833. 26, 38

Liu, J., Akhtar, N. & Mian, A. (2017c). Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition. *arXiv preprint arXiv:1711.05941, 2017*. xi, 23, 24

Liu, J., Wang, G., Duan, L.Y., Abdiyeva, K. & Kot, A.C. (2018). Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Trans. Image Process.*, **27**, 1586–1599. 26, 38

LIU, M. & LIU, H. (2016). Depth context: a new descriptor for human activity recognition by using sole depth sequences. *Neurocomputing*, **175**, 747–758. 15, 37

LIU, M., LIU, H. & CHEN, C. (2017d). Robust 3d action recognition through sampling local appearances and global distributions. *IEEE Trans. Multimedia*. 37, 81

LIU, Z., ZHANG, C. & TIAN, Y. (2016c). 3d-based deep convolutional neural network for action recognition with depth sequences. *Image Vis. Comput.*, **55**, 93–100. 28, 37

LUO, Z., PENG, B., HUANG, D.A., ALAHI, A. & FEI-FEI, L. (2017). Unsupervised learning of long-term motion dynamics for videos. In *Proc. Conf. Comput. Vis. Pattern Recog.*. 37

MAHASSENI, B. & TODOROVIC, S. (2016). Regularizing long short term memory with 3d human-skeleton sequences for action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 3054–3062. 25

MIAO, Q., LI, Y., OUYANG, W., MA, Z., XU, X., SHI, W., CAO, X., LIU, Z., CHAI, X., LIU, Z. *et al.* (2017). Multimodal gesture recognition based on the resc3d network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 3047–3055. 27

MITTAL, A. & PARAGIOS, N. (2004). Motion-based background subtraction using adaptive kernel density estimation. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, II–II, Ieee. 71

MOLCHANOV, P., YANG, X., GUPTA, S., KIM, K., TYREE, S. & KAUTZ, J. (2016). Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proc. Conf. Comput. Vis. Pattern Recog.*, 4207–4215. 30

MÜLLER, M., RÖDER, T., CLAUSEN, M., EBERHARDT, B., KRÜGER, B. & WEBER, A. (2007). Documentation mocap database hdm05. xi, 8, 9

NI, B., PEI, Y., MOULIN, P. & YAN, S. (2013). Multilevel depth and image fusion for human activity detection. *IEEE Trans. Cybern.*, **43**, 1383–1394. 22, 56

NIEBLES, J.C. & FEI-FEI, L. (2007). A hierarchical model of shape and appearance for human action classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1–8. 40

NIEBLES, J.C., WANG, H. & FEI-FEI, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vis.*, **79**, 299–318. 40

NOWOZIN, S. & SHOTTON, J. (2012). Action points: A representation for low-latency online human action recognition. *Microsoft Research Cambridge, Tech. Rep. MSR-TR-2012-68*. 30

NÚÑEZ, J.C., CABIDO, R., PANTRIGO, J.J., MONTEMAYOR, A.S. & VÉLEZ, J.F. (2018). Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition. *Pattern Recog.*, **76**, 80–94. 26, 37

OFLI, F., CHAUDHRY, R., KURILLO, G., VIDAL, R. & BAJCSY, R. (2014). Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *J. Vis. Commun. Image Represent.*, **25**, 24–38. 12, 13

OHN-BAR, E. & TRIVEDI, M.M. (2013). Joint angles similarities and hog2 for action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 465–470. 12, 37

OREIFEJ, O. & LIU, Z. (2013). Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 716–723. 15, 37, 53, 54, 81

PATRON-PEREZ, A., MARSZALEK, M., ZISSERMAN, A. & REID, I.D. (2010). High five: Recognising human interactions in tv shows. *BMVC*, **1**, 2. 33

PAZHOUMAND-DAR, H., LAM, C.P. & MASEK, M. (2015). Joint movement similarities for robust 3d action recognition using skeletal data. *J. Vis. Commun. Image Represent.*, **30**, 10–21. 13, 40

PRESTI, L.L. & LA CASCIA, M. (2016). 3d skeleton-based human action classification: A survey. *Pattern Recognition*, **53**, 130–147. 11

QIAO, R., LIU, L., SHEN, C. & VAN DEN HENGEL, A. (2017). Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition. *Pattern Recog.*, **66**, 202–212. 12, 37, 40, 81

RAHMANI, H. & BENNAMOUN, M. (2017). Learning action recognition model from depth and skeleton videos. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 5832–5841. 27

RAHMANI, H. & MIAN, A. (2016). 3d action recognition from novel viewpoints. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1506–1515. 27

RAMAN, N. & MAYBANK, S. (2016). Activity recognition using a supervised non-parametric hierarchical hmm. *Neurocomputing*, **199**, 163–177. 16

RYOO, M. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proc. IEEE Int. Conf. Comput. Vis.*, 1036–1043. 70, 84

RYOO, M.S. & AGGARWAL, J.K. (2010). UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). 33

SAHA, S., KONAR, A. & JANARTHANAN, R. (2015). Two person interaction detection using kinect sensor. *Facets of Uncertainties and Applicat.*, 167–176. 20

SEIDENARI, L., VARANO, V., BERRETTI, S., DEL BIMBO, A. & PALA, P. (2013). Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 479–485. 6, 31, 54, 55

SHAHROUDY, A., LIU, J., NG, T.T. & WANG, G. (2016a). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1010–1019. 24, 25, 34

SHAHROUDY, A., NG, T.T., YANG, Q. & WANG, G. (2016b). Multimodal multipart learning for action recognition in depth videos. *IEEE Trans. Pattern Anal. and Mach. Intell.*, **38**, 2123–2129. 3, 16, 37

SHAHROUDY, A., NG, T.T., YANG, Q. & WANG, G. (2016c). Multimodal multipart learning for action recognition in depth videos. 70

SHAN, J. & AKELLA, S. (2014). 3d human action segmentation and recognition using pose kinetic energy. In *IEEE Workshop Advanced Robotics and its Social Impacts*, 69–75. 4, 29

SHARAF, A., TORKI, M., HUSSEIN, M.E. & EL-SABAN, M. (2015). Real-time multi-scale action detection from 3d skeleton data. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, 998–1005. 30

SHI, Y., TIAN, Y., WANG, Y. & HUANG, T. (2017). Sequential deep trajectory descriptor for action recognition with three-stream cnn. *IEEE Trans. Multimedia*, **19**, 1510–1520. 81

SHI, Z. & KIM, T.K. (2017). Learning and refining of privileged information-based rnns for action recognition from depth sequences. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*. 28, 37

SHOTTON, J., SHARP, T., KIPMAN, A., FITZGIBBON, A., FINOCCHIO, M., BLAKE, A., COOK, M. & MOORE, R. (2013a). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, **56**, 116–124. xi, 2, 10

SHOTTON, J., SHARP, T., KIPMAN, A., FITZGIBBON, A., FINOCCHIO, M., BLAKE, A., COOK, M. & MOORE, R. (2013b). Real-time human pose recognition in parts from single depth images. *Commun. ACM*, **56**, 116–124. 12, 40, 85

SHOU, Z., WANG, D. & CHANG, S. (2016). Action temporal localization in untrimmed videos via multi-stage cnns. In *Proc. Conf. Comput. Vis. Pattern Recog.*, vol. 3. 31

SIMONYAN, K. & ZISSERMAN, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Representations*. 23

SINGHAL, A. *et al.* (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, **24**, 35–43. 45

SLAMA, R., WANNOUS, H. & DAOUDI, M. (2014). Grassmannian representation of motion depth for 3d human gesture and action recognition. In *Proc. Int. Conf. Pattern Recog.*, 3499–3504. 15

SONG, S., LAN, C., XING, J., ZENG, W. & LIU, J. (2017). An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 4263–4270. 26, 38, 67

SONG, S., LAN, C., XING, J., ZENG, W. & LIU, J. (2018). Spatio-temporal attention based lstm networks for 3d action recognition and detection. *IEEE Trans. Image Process.*, **pp**, 1–1. 31

SONG, Y., DEMIRDJIAN, D. & DAVIS, R. (2012). Continuous body and hand gesture recognition for natural human-computer interaction. *ACM Trans. Ineractive Intell. Syst.*, **2**, 5. 30

SUNG, J., PONCE, C., SELMAN, B. & SAXENA, A. (2012). Unstructured human activity detection from rgbd images. In *Proc. IEEE Int. Conf. Robotics and Automation*, 842–849. 17

SWAIN, M.J. & BALLARD, D.H. (1991). Color indexing. *Int. J. Comput. Vis.*, **7**, 11–32. 59

THEODORAKOPOULOS, I., KASTANIOTIS, D., ECONOMOU, G. & FOTOPOULOS, S. (2014). Pose-based human action recognition via sparse representation in dissimilarity space. *J. Vis. Commun. Image Represent.*, **25**, 12–23. 13, 37

TOSHEV, A. & SZEGEDY, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1653–1660. 12

TRABELSI, R., VARADARAJAN, J., PEI, Y., ZHANG, L., JABRI, I., BOUALLEGUE, A. & MOULIN, P. (2017). Robust multi-modal cues for dyadic human interaction recognition. In *Proc. Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*, 47–53. 21

TRAN, D., BOURDEV, L., FERGUS, R., TORRESANI, L. & PALURI, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, 4489–4497. 28

van Gemeren, C., Tan, R.T., Poppe, R. & Veltkamp, R.C. (2014). Dyadic interaction detection from pose and flow. *Human Behavior Understanding*, 101–115. 21, 64

van Gemeren, C., Poppe, R. & Veltkamp, R.C. (2016). Spatio-temporal detection of fine-grained dyadic human interactions. In *Int. Workshop on Human Behavior Understanding*, 116–133. 64

Van Gemert, J.C., Veenman, C.J., Smeulders, A.W. & Geusebroek, J.M. (2010). Visual word ambiguity. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**, 1271–1283. 45

Veeriah, V., Zhuang, N. & Qi, G.J. (2015). Differential recurrent neural networks for action recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, 4041–4049. 25, 37

Vemulapalli, R. & Chellapa, R. (2016). Rolling rotations for recognizing human actions from 3d skeletal data. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 4471–4479. 24

Vemulapalli, R., Arrate, F. & Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 588–595. 3, 13, 37, 47, 50, 53, 54, 55

Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z. & Campos, M.F. (2014). On the improvement of human action recognition from depth map sequences using space–time occupancy patterns. *Pattern Recog. Lett.*, **36**, 221–227. 52, 53, 54

Wang, C., Wang, Y. & Yuille, A.L. (2013). An approach to pose-based action recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 915–922. 53, 54

Wang, C., Liu, Z. & Chan, S.C. (2015a). Superpixel-based hand gesture recognition with kinect depth camera. *IEEE Trans. Multimedia*, **17**, 29–39. 81

Wang, C., Wang, Y. & Yuille, A.L. (2016a). Mining 3d key-pose-motifs for action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2639–2647. 37

WANG, H. & WANG, L. (2017). Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks. *Proc. Conf. Comput. Vis. Pattern Recog.*, 499–508. 26, 38

WANG, H. & WANG, L. (2018). Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Trans. Image Process.*, **27**, 4382–4394. 92

WANG, J., LIU, Z., CHOROWSKI, J., CHEN, Z. & WU, Y. (2012a). Robust 3d action recognition with random occupancy patterns. In *Proc. Eur. Conf. Comput. Vis.*, 872–885. 15, 53, 54

WANG, J., LIU, Z., WU, Y. & YUAN, J. (2012b). Mining actionlet ensemble for action recognition with depth cameras. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1290–1297. xi, 16, 32, 37, 53, 54, 57

WANG, J., LIU, Z., WU, Y. & YUAN, J. (2014). Learning actionlet ensemble for 3d human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell*, **36**, 914–927. 16, 37

WANG, P., LI, W., GAO, Z., TANG, C., ZHANG, J. & OGUNBONA, P. (2015b). Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring. In *Proc. 23rd ACM int. conf. Multimedia*, 1119–1122. 27

WANG, P., LI, W., GAO, Z., ZHANG, J., TANG, C. & OGUNBONA, P. (2015c). Deep convolutional neural networks for action recognition using depth map sequences. *arXiv preprint arXiv:1501.04686, 2015*. 37

WANG, P., LI, W., GAO, Z., ZHANG, J., TANG, C. & OGUNBONA, P.O. (2015d). Action recognition from depth maps using deep convolutional neural networks. 14

WANG, P., LI, W., GAO, Z., ZHANG, J., TANG, C. & OGUNBONA, P.O. (2016b). Action recognition from depth maps using deep convolutional neural networks. *IEEE Trans. Human Mach. Syst.*, **46**, 498–509. 24, 27, 36, 37

WANG, P., LI, W., LIU, S., GAO, Z., TANG, C. & OGUNBONA, P. (2016c). Large-scale isolated gesture recognition using convolutional neural networks. In *Int. Conf. Pattern Recog.*, 7–12. 27

WANG, P., LI, W., GAO, Z., ZHANG, Y., TANG, C. & OGUNBONA, P. (2017). Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*. 28

WANG, X., GAO, L., WANG, P., SUN, X. & LIU, X. (2018). Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. *IEEE Transactions on Multimedia*, **20**, 634–644. 81

WEINLAND, D., RONFARD, R. & BOYER, E. (2006). Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Understanding*, **104**, 249–257. 23

WU, C., ZHANG, J., SAVARESE, S. & SAXENA, A. (2015). Watch-n-patch: Unsupervised understanding of actions and relations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 4362–4370. 29, 30

WU, D., PIGOU, L., KINDERMANS, P.J., LE, N.D.H., SHAO, L., DAMBRE, J. & ODOBEZ, J.M. (2016a). Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **38**, 1583–1597. 27, 28

WU, H., SHAO, J., XU, X., JI, Y., SHEN, F. & SHEN, H.T. (2017). Recognition and detection of two-person interactive actions using automatically selected skeleton features. *IEEE Trans. Human Mach. Syst., 2017*. 20, 30, 38

WU, M.Y., CHEN, T.Y., CHEN, K.Y. & FU, L.C. (2016b). Daily activity recognition using the informative features from skeletal and depth data. In *Proc. IEEE Int. Conf. Robotics and Automation*, 1628–1633. 37

XIA, L. & AGGARWAL, J. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2834–2841. 15, 37, 53, 54

XIA, L., CHEN, C.C. & AGGARWAL, J. (2012). View invariant human action recognition using histograms of 3d joints. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 20–27. 3, 13, 52, 53, 54

XIA, L., GORI, I., AGGARWAL, J. & RYOO, M. (2015). Robot-centric activity recognition from first-person rgb-d videos. In *IEEE Winter Conf. Applicat. Comput. Vis.*, 357–364. 21

XU, N., LIU, A., NIE, W., WONG, Y., LI, F. & SU, Y. (2015). Multi-modal & multi-view & interactive benchmark dataset for human action recognition. In *Proc. ACM Int. Conf. Multimedia*, 1195–1198. 21

YAN, S., XIONG, Y. & LIN, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455, 2018*. 24

YANG, X. & TIAN, Y. (2012). Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 14–19. xi, 17, 47, 52, 53, 54

YANG, X. & TIAN, Y. (2014a). Effective 3d action recognition using eigenjoints. *J. Vis. Commun. Image Represent.*, **25**, 2–11. 16

YANG, X. & TIAN, Y. (2014b). Super normal vector for activity recognition using depth sequences. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 804–811. 15, 37, 53, 54

YANG, X., ZHANG, C. & TIAN, Y. (2012a). Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proc. ACM Int. Conf. Multimedia*, 1057–1060. xi, 14, 27

YANG, Y. & RAMANAN, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell*, **35**, 2878–2890. 12

YANG, Y., BAKER, S., KANNAN, A. & RAMANAN, D. (2012b). Recognizing proxemics in personal photos. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 3522–3529. 52, 53, 54

YANG, Y., DENG, C., GAO, S., LIU, W., TAO, D. & GAO, X. (2017a). Discriminative multi-instance multitask learning for 3d action recognition. *IEEE Trans. Multimedia*, **19**, 519–529. 53, 54

YANG, Y., DENG, C., TAO, D., ZHANG, S., LIU, W. & GAO, X. (2017b). Latent max-margin multitask learning with skelets for 3-d action recognition. *IEEE Trans. Cybern.*, **47**, 439–448. 53, 54

YE, M., WANG, X., YANG, R., REN, L. & POLLEFEYS, M. (2011). Accurate 3d pose estimation from a single depth image. In *Proc. IEEE Int. Conf. Comput. Vis.*, 731–738. 12

YUN, K., HONORIO, J., CHATTOPADHYAY, D., BERG, T.L. & SAMARAS, D. (2012). Two-person interaction detection using body-pose features and multiple instance learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 28–35. xi, 4, 5, 6, 19, 20, 21, 33, 38, 56, 64, 65, 67, 68

ZANFIR, M., LEORDEANU, M. & SMINCHISESCU, C. (2013). The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2752–2759. 3, 12, 37, 50, 53, 54, 81

ZHANG, H. & PARKER, L.E. (2016). Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos. *IEEE Trans. Circuits Syst. Video Technol.*, **26**, 541–555. 17, 37

ZHANG, J., LI, W., OGUNBONA, P.O., WANG, P. & TANG, C. (2016). Rgb-d-based action recognition datasets: A survey. *Pattern Recog.*, **60**, 86–105. 11

ZHANG, L., ZHU, G., SHEN, P., SONG, J., SHAH, S.A. & BENNAMOUN, M. (2017a). Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 3120–3128. 27

ZHANG, P., LAN, C., XING, J., ZENG, W., XUE, J. & ZHENG, N. (2017b). View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2117–2126. 26, 38

ZHANG, S., LIU, X. & XIAO, J. (2017c). On geometric features for skeleton-based action recognition using multilayer lstm networks. In *IEEE Winter Conf. Applications Comput. Vis.*, 148–157. 25

ZHANG, S., YANG, Y., XIAO, J., LIU, X., YANG, Y., XIE, D. & ZHUANG, Y. (2018). Fusing geometric features for skeleton-based action recognition using multilayer lstm networks. *IEEE Trans. Multimedia*. 81

ZHU, F., SHAO, L., XIE, J. & FANG, Y. (2016a). From handcrafted to learned representations for human action recognition: a survey. *Image and Vis. Comput., 2016*. 11

ZHU, G., ZHANG, L., SHEN, P. & SONG, J. (2016b). An online continuous human action recognition algorithm based on the kinect sensor. *Sensors*, **16**, 161. 4, 29

ZHU, W., LAN, C., XING, J., ZENG, W., LI, Y., SHEN, L. & XIE, X. (2016c). Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks. In *AAAI*, 3697–3703. xi, 25, 38

ZHU, W., LAN, C., XING, J., ZENG, W., LI, Y., SHEN, L. & XIE, X. (2016d). Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. *AAAI*. 67

ZHU, Y., CHEN, W. & GUO, G. (2013). Fusing spatiotemporal features and joints for 3d action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 486–491. 16, 37

Appendix A

Publications

A.1 Journal Papers

1. Liu, B., Cai, H., Ju, Z., and Liu, H. (2018). RGB-D Sensing based Human Action and Interaction Analysis: A Survey, *Pattern Recognition*,(**Under Review**).

2. Liu, B., Ju, Z., Cai, H., and Liu, H. (2018). Multi-stage Soft Regression for Online Activity Recognition, *IEEE Transactions on Multimedia*,(**Under Review**).

3. Liu, B., Ju, Z., and Liu, H. (2018). A Structured Multi-Feature Representation for Recognizing Human Action and Interaction, *Neurocomputing*, (**Accepted**).

4. Cai, H., Liu, B., Zhang, J., Chen, S. and Liu, H. (2017). Visual focus of attention estimation using eye center localization, *IEEE Systems Journal*

A.2 Conference Papers

1. Liu, B., Ju, Z., Kubota, N., Liu,H. (2018). Online Action Recognition based on Skeleton Motion Distribution. *British Machine Vision Conference Workshop*, (**Accepted**).

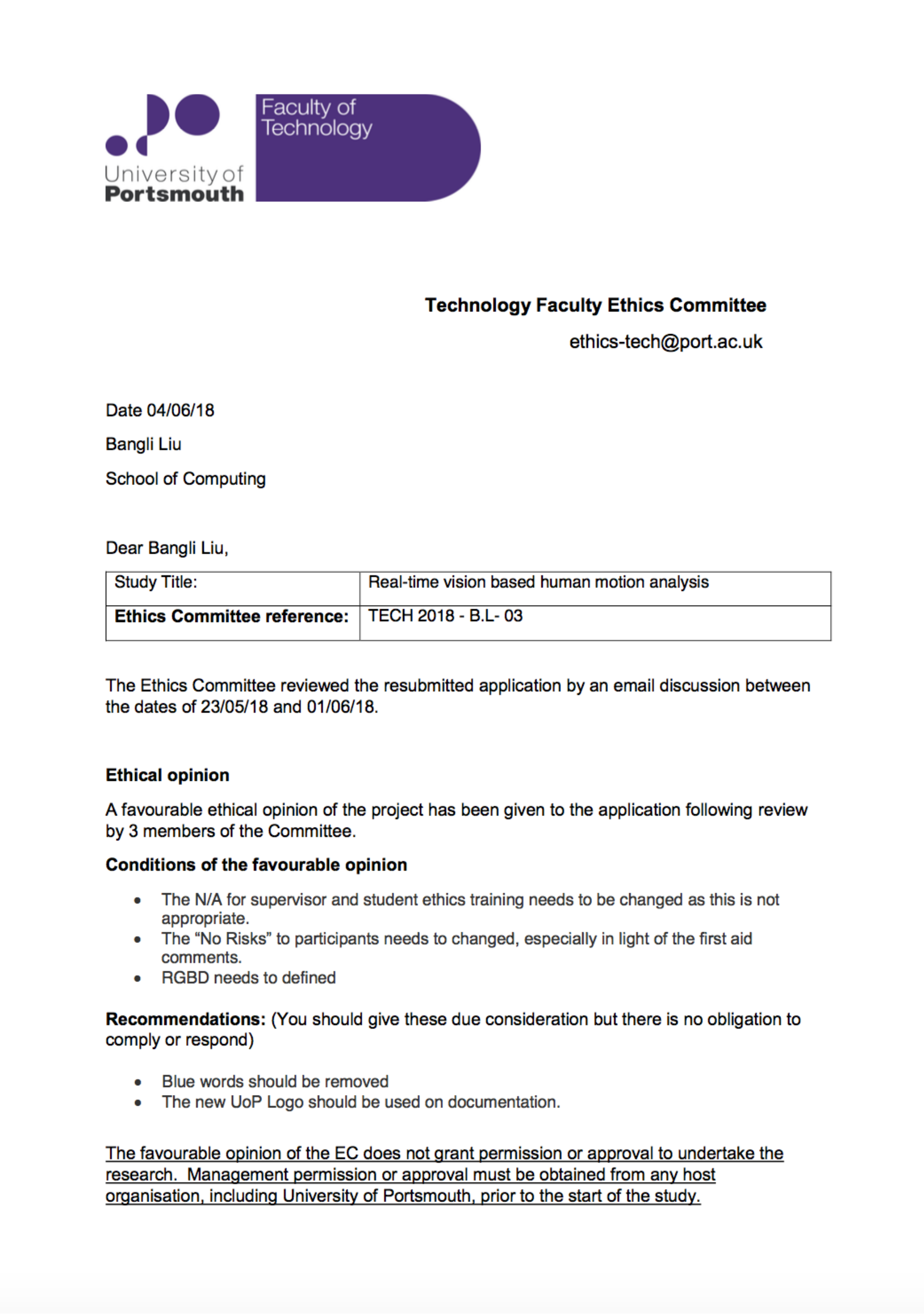
2. Cai, H., Liu, B., Ju, Z., Thill, S., Belpaeme, T., Vanderborght, B., Liu,H.(2018). Accurate eye center localization via hierarchical adaptive convolution. *British Machine Vision Conference*, (**Accepted**).

3. Liu, B., Cai, H., Ji, X., Liu,H. (2017). Human-human interaction recognition based on spatial and motion trend feature. *Int. Conf. Image Processing (ICIP)*, 4547-4551.

4. Liu, B., Yu, H., Zhou, X., Liu, H. (2016). Combining 3D joints Moving Trend and Geometry property for human action recognition. *IEEE Int. Conf. Systems, Man, and Cybernetics*.

Appendix B

Research Ethics



University of Portsmouth

Faculty of Technology

Technology Faculty Ethics Committee

ethics-tech@port.ac.uk

Date 04/06/18

Bangli Liu

School of Computing

Dear Bangli Liu,

Study Title:	Real-time vision based human motion analysis
Ethics Committee reference:	TECH 2018 - B.L- 03

The Ethics Committee reviewed the resubmitted application by an email discussion between the dates of 23/05/18 and 01/06/18.

Ethical opinion

A favourable ethical opinion of the project has been given to the application following review by 3 members of the Committee.

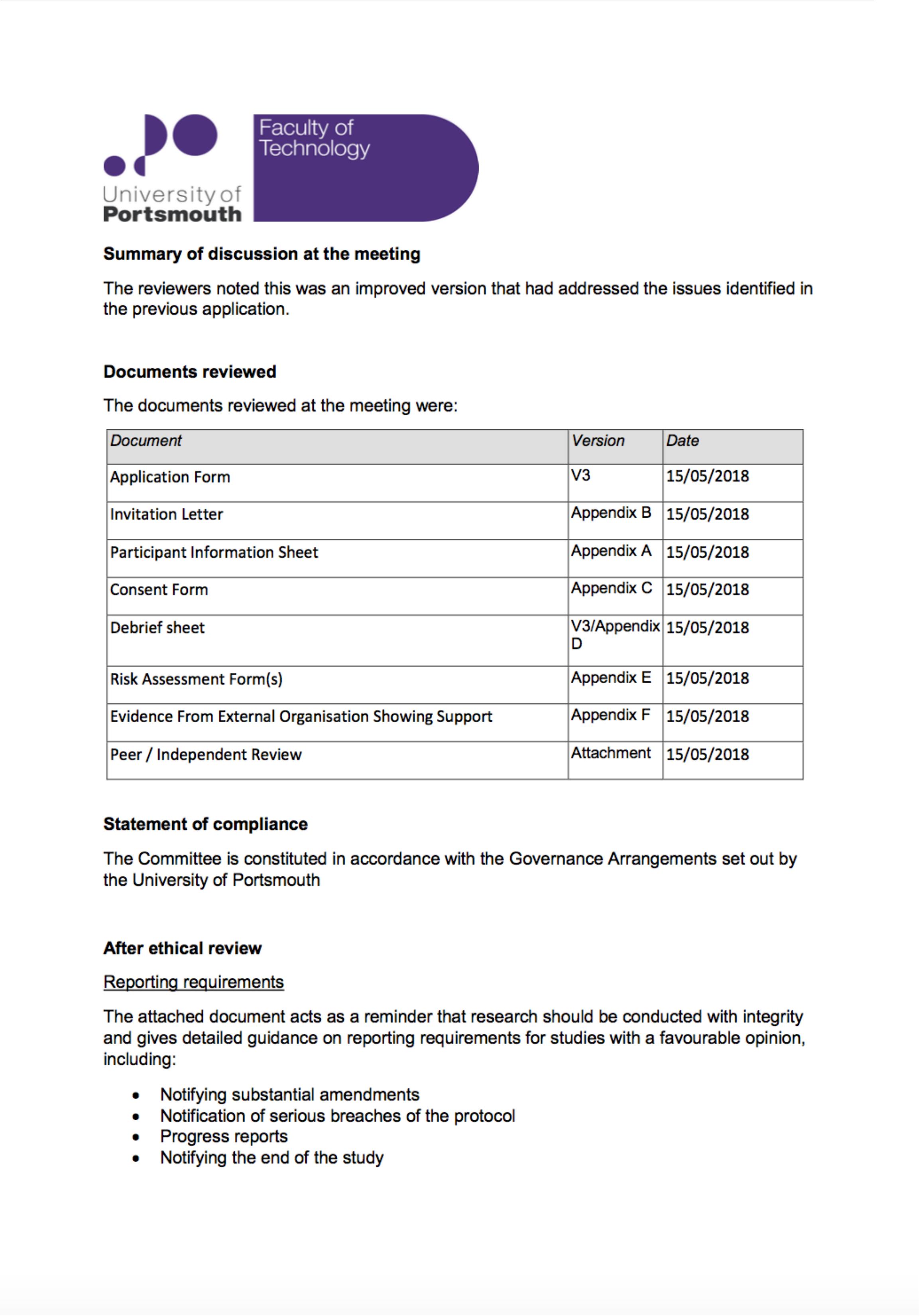
Conditions of the favourable opinion

- The N/A for supervisor and student ethics training needs to be changed as this is not appropriate.
- The “No Risks” to participants needs to changed, especially in light of the first aid comments.
- RGBD needs to defined

Recommendations: (You should give these due consideration but there is no obligation to comply or respond)

- Blue words should be removed
- The new UoP Logo should be used on documentation.

The favourable opinion of the EC does not grant permission or approval to undertake the research. Management permission or approval must be obtained from any host organisation, including University of Portsmouth, prior to the start of the study.



University of Portsmouth

Faculty of Technology

Summary of discussion at the meeting

The reviewers noted this was an improved version that had addressed the issues identified in the previous application.

Documents reviewed

The documents reviewed at the meeting were:

Document	*Version*	*Date*
Application Form	V3	15/05/2018
Invitation Letter	Appendix B	15/05/2018
Participant Information Sheet	Appendix A	15/05/2018
Consent Form	Appendix C	15/05/2018
Debrief sheet	V3/Appendix D	15/05/2018
Risk Assessment Form(s)	Appendix E	15/05/2018
Evidence From External Organisation Showing Support	Appendix F	15/05/2018
Peer / Independent Review	Attachment	15/05/2018

Statement of compliance

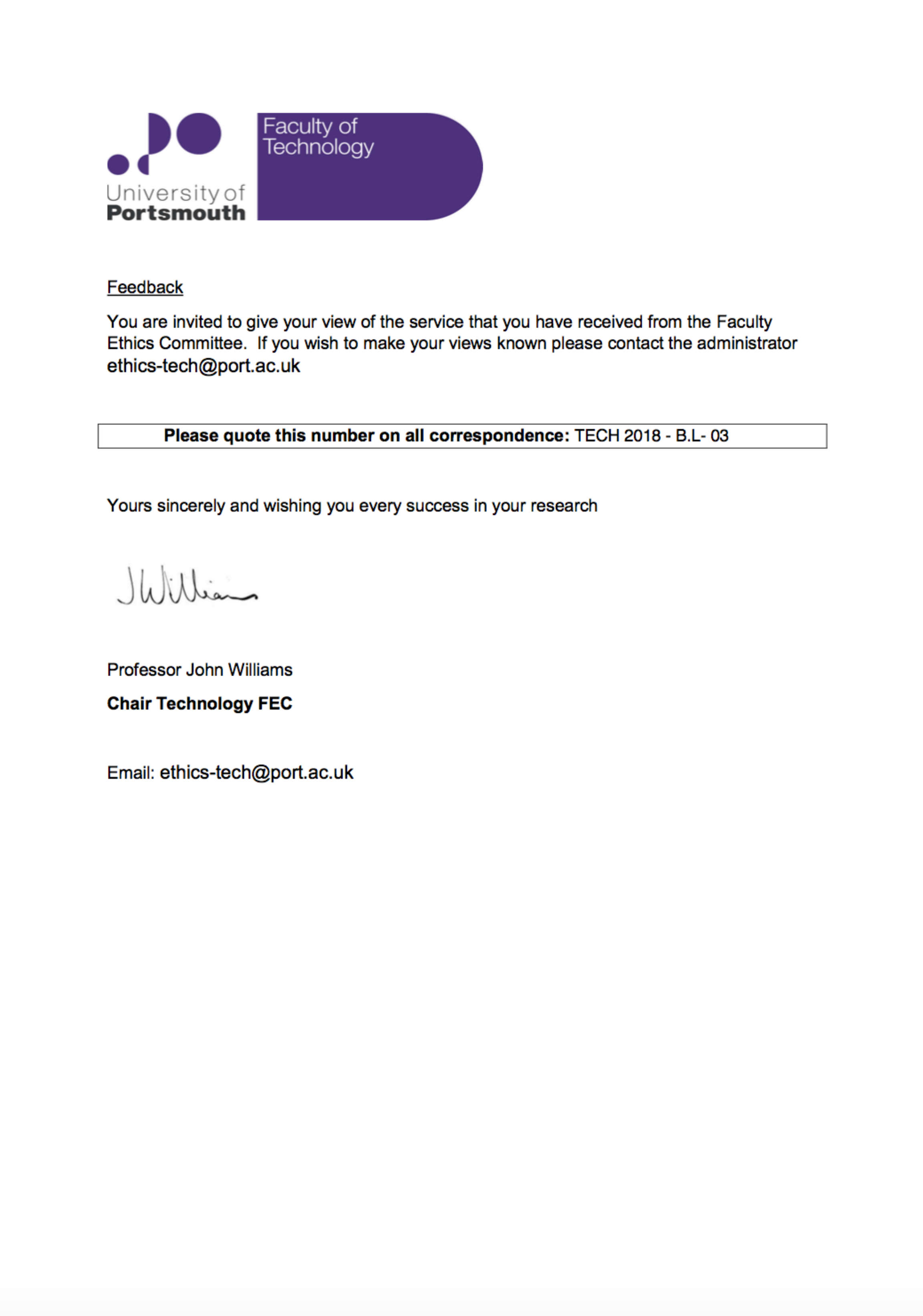
The Committee is constituted in accordance with the Governance Arrangements set out by the University of Portsmouth

After ethical review

Reporting requirements

The attached document acts as a reminder that research should be conducted with integrity and gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Notification of serious breaches of the protocol
- Progress reports
- Notifying the end of the study



University of Portsmouth

Faculty of Technology

Feedback

You are invited to give your view of the service that you have received from the Faculty Ethics Committee. If you wish to make your views known please contact the administrator ethics-tech@port.ac.uk

Please quote this number on all correspondence: TECH 2018 - B.L- 03

Yours sincerely and wishing you every success in your research

JWilliams

Professor John Williams

Chair Technology FEC

Email: ethics-tech@port.ac.uk

FORM UPR16

Research Ethics Review Checklist

University of Portsmouth

Please include this completed form as an appendix to your thesis (see the Postgraduate Research Student Handbook for more information

Postgraduate Research Student (PGRS) Information		**Student ID:**	797995
PGRS Name:	Bangli Liu		
Department:	School of Computing	**First Supervisor:**	Honghai Liu
Start Date: (or progression date for Prof Doc students)	**1 October 2015**		
Study Mode and Route:	Part-time □ Full-time ☒	MPhil □ PhD ☒	MD □ Professional Doctorate □

Title of Thesis:	Vision-based Human Activity Analysis
Thesis Word Count: (excluding ancillary data)	33571

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES ☒ NO □
b) Have all contributions to knowledge been acknowledged?	YES ☒ NO □
c) Have you complied with all agreements relating to intellectual property, publication and authorship?	YES ☒ NO □
d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES ☒ NO □
e) Does your research comply with all legal, ethical, and contractual requirements?	YES ☒ NO □

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	TECH 2018 - B.L- 03

If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:

Signed *(PGRS)*:	**Date:** 31/08/2018

	Bang li Lin	