UNIVERSITÉ DE CAEN NORMANDIE

U.F.R. de Sciences

ÉCOLE DOCTORALE SIMEM

THÈSE

Présentée par

M. Adnan Salih Sahle AL ALWANI

soutenue le

xx xx 2016

en vue de l'obtention du

DOCTORAT de l'UNIVERSITÉ de CAEN

Spécialité : Informatique et applications

Arrêté du 07 août 2006

Event and action recognition from thermal and 3D depth Sensing

Laboratoire : Groupe de recherche en informatique, image, automatique et instrumentation de Caen (GREYC)

Rapporteurs	Pr Abdelmalik Taleb-Ahmed, LAMIH, UMR CNRS 8201, UVHC Pr Charles Tijus, LUTIN -CHART, Paris 8
Examinateurs	Pr Alain Bretto, GREYC, UMR CNRS UMR 6072, Caen Pr François Jouen, CHART- EPHE, Paris 8 Pr Luigi Lancieri, CRISTAL, UMR CNRS 8219 , Lille MdC-HDR Youssef Chahir, GREYC, UMR CNRS UMR 6072, Caen

Abstract

Modern computer vision algorithms try to understand the human activity using 3D visible sensors. However, there are inherent problems using 2D visible sensors as a data source. First, visible light images are sensitive to illumination changes and background clutter. Second, the 3D structural information of the scene is degraded when mapping the 3D scene to 2D images. Recently, the easy access to the RGBD data at real-time frame rate is leading to a revolution in perception and inspired many new research. Time of Flight (ToF) and multi-view sensors have been used to model the 3D structure of the scene.

Otherwise, infrared thermography (IRT), also known as thermal imaging, is an ideal technology to investigate thermal anomalie under different circumstances because it provides complete thermal images of an object with no physical attachments (nonintrusive). IRT is now being introduced to a wide range of different applications, such as medical diagnostic and surveillance.

However, finding meaningful features from a time series data from thermal video is still a challenging problem, especially for event detection. This problem is particularly hard due to enormous variations in visual and motion appearance of object, moving background, occlusions and thermal noise.

In this thesis, we propose a framework for the detection of visual events in thermal video and 3d human actions in RGBD data. Despite differences in the applications, the associated fundamental problems share numerous properties, for instance the necessity of handling vision-based approach for the automatic recognition of events.

□ The first part of the thesis deals with the recognition of events in thermal video. In this context, the use of time series is challenging due to the graphical nature which exposes hidden patterns and structural changes in data. In this study, we investigated the use of visual texture patterns for time series classification. Our principal aim was to develop a general framework for time series data mining based on event analysis with an application to the medical domain. In particular, we are interested to pain/no-pain detection using parametric statistics and shape descriptors in order to analyze and to classify time 2D distribution data sets.

We first extracted automatically thermal-visual facial features from each face considered as the region of interest (ROI) of the image.

We proposed two feature descriptors for the signal pattern of interest (POI) which efficiently exploits the dependence between time and frequency in one-dimension (1D) signal. The original signal is extracted directly from local patch in ROI.
The first method is based on ***non-redundant temporal local binary pattern*** (NRTLBP).
The second approach propose a ***topological persistence descriptor*** (TP) for extracting and filtering local extrema of 1D signal. Local minima and local maxima are extracted, paired, and sorted according to their persistence.
The final representation of an event is a completely new feature vector of all paired critical values. These features provide many benefits for many applications to get a fast estimation of the event in dynamic time series data.
Both methods are validated using an Extreme Learning Machine (ELM) and Support vector Machine (SVM) classifiers.
Experimental results on a real thermal-based data set ”Pain in Preterm Infants” (PPI), which is captured in a real condition monitoring environment, show that the proposed methods successfully capture temporal changes in events and achieve higher recognition rates. PPI dataset was developed in the context of Infant pain project, a french project supported by the French National Research Agency Projects for science (ANR).

□ In the second part of the thesis, we investigate the problem of recognizing human activities in different application scenarios: controlled video environment(e.g. indoor surveillance) and specially depth or skeletal data (e.g. captured by Kinect). We focus on developing spatio-temporal features, and applying these features to identify human activities from a sequence of RGB-D images, i.e.,color images with depth information.

First, we proposed a view-invariant approach which use ***joint angles and relative joint positions as features***. These features are quantized into posture visual words and their temporal transitions are encoded as observation symbols in a Hidden Markov Model (HMM). To eliminate rotation dependence in skeletal descriptors, we proposed an approach that combines the ***covariance descriptor and the spherical harmonics*** (SHs). The harmonic representation of 3d shape descriptors is adapted to skeleton joint-based human action recognition. To improve the accuracy and the convergence speed of the SHs solutions, we proposed an extension of the model, using ***quadratic spherical harmonics*** (QSH) representation, to encode pose information in the spatiotemporal space. These SHs representations are compact and discriminating. For the recognition task, we used ELM classifier. Our experimental results on a number of popular 3d action datasets show significant achievements in terms of accuracy, scalability and efficiency in comparison to alternate methods, of the state-of-the-art.

Résumé

Acknowledgements

Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

Contents

1.1 Overview

1.1.1 Thermal Sensing

Thermal imaging has become the first choice today for diagnostic imaging in general practice. This technology is changing the practice of the most serious societal

problems. Until now, most of the vision algorithms and physiological events assessment were built on thermal imaging. However, thermal signature is important for many specific vision applications such as condition monitoring, dark object search , and it may bring significant improvement to the current vision tasks including health diagnostic, scene recognition under different lighting conditions, medical analysis, legal and insurance industries.
Thermal imaging technology does not replace or diminish the value of existing visible-bases vision methods. They too are advancing in quality and effectiveness. Thermal imaging simply provides access into earlier and safer diagnosis for many complex cases and degenerative conditions.

Thermal imaging is so safe and effective that it is becoming of interest to alleviate some problems attached to standard imaging techniques. In a thermal image that consists of objects in a scene, object structure can be easily extracted from the background regardless of lighting conditions and colors of the foreground surfaces and backgrounds, because the temperatures of the human body and background are different in most situations. In other side, thermal image analysis can be investigated to detect at a distance facial patterns of anxiety, alertness, which will help a specialist distinguish pain patterns as physical or psychogenic, or even reveal underlying pain issues.
The rapid progressive development of infrared cameras (passive sensor), give us easy access to the thermal signature data at a higher resolution. The easy access to real-time thermal data is simply provides access into safer diagnosis of temperature distribution, and leading to a revolution in activity and event recognition task.

Among the many aspects of computer vision, the problem of event recognition in thermal videos has become an increasingly popular due to its demand and applications in a range of areas. Some of its application areas are automated condition monitoring systems, health-care diagnostic and monitoring, and human activity recognition. The task of event analysis in thermal video involve identifying the temporal range of an event in a video and sometimes the location of the event. While there have been increasing efforts recently to tackle this problem specially in premature assessment application, it remains rather challenging due to compounding issues such as large variation of thermal signature, varied durations of events, and noise. To the best of our knowledge, there was no literature addressing the problem of real event recognition in daily monitoring care system and thermal data.

Thermal assistive system equipped with event recognition algorithms can describe the vital signal to people with visual impairments. Moreover, event prediction techniques could assist remote monitoring care by reporting abnormal behavior of objects to health care workers.

We propose to take advantage of this readily available thermal assistive technique to improve the performance of event recognition algorithms. Especially, we address the behavioral responses to pain in Condition Monitoring Neonatal Intensive Care CMNIC using thermal sensing.

1.1.2 3D imaging

During the last decades vision-based human action recognition has been focused on traditional 2D imaging technique. It has contributed to the solution of some of the most topic in the computer vision research. However, 3D data is important to overcome the limitations with 2D vision, such as shadow, view-invariant, occlusion. This 3D structure may achieve significant improvement to the recent computer vision fields, such as activity analysis, human computer interaction, action recognition. and robot vision.
The acquisition of 3D imagery is basically addressed in two techniques. The first one is to reconstruct the 3D geometry from 2D multiple camera, upon this imaging scheme. 3D reconstruction is limited to the overlapping area of the camera views, leading to the difficulty understanding its 3D structure. Thus, there might be valuable information embedded in the 2D image to infer 3D. However, it is extremely challenging for current computer vision algorithms to reconstruct 3D from 2D images, due to the problem of the information losses when mapping the 3D geometry into a 2D image. The second way is to model 3D structure directly from Time of Flight ToF range cameras. Particularly, the first generation of range sensors were either too expensive, difficult to use in indoor environments, slow capturing rate, or provided poor estimation of distance.

The development of 3D imaging sensor has progressed rapidly over the past decade. Recently, the release of the RGB-D cameras at relatively low costs and easily handle, gives computer vision researchers an easy access to the 3D structural information at a higher frame rate and resolution. The invention of the real time RGBD imaging have directed a vision techniques into consumers living rooms, and is leading to a revolution in computer vision applications. These systems provide multimodal data (2D image,depth sound, skeleton) that offer a rich perception of the environment and human activities. Moreover, RGBD-based 3D sensing is provided the solution to many invariances. Such as simplifying the segmentation of objects, recognize activities in complex scenes and the detection of occlusions. Combining the advantages of low-cost and real-time nature of the RGBD devices. The quality of the depth imaging is competitive, even their are imperfection issues such the noise, lack of data precision, and occlusion.

However, we propose to take advantage of the availability of RGBD-based human structure, to contribute with existing computer vision algorithms to improve skeleton-based action recognition methods. Especially, much of the work presented in this thesis is proposed to address the problem of human action recognition based on information from depth channel. We show that even under the many constraints of the quality levels of the RGBD images, the achievements made by the 3D skeleton structure are still quite encouraging. However, we believe that the rapid generation of the RGBD sensors will provides better quality depth map with accurate body articulated data. Which has been made available to the community to support future work.

1.1.3 3D Human body configuration and pose estimation

The use of 3D data allows for efficient analysis of 3D human actions. While 2D human action recognition has received high interest during the last decade, 3D human action recognition is still a less explored field. Relatively few authors have so far reported work on human motion capture, body modeling, pose estimation and action recognition in more general [Ziming et al., 2008, Moeslund et al., 2006, Poppe, 2010, Weinland et al., 2011]. In recent years, a wide range of applications using 3D human body modeling, pose estimation and activity recognition has been introduced. To comply with the requirements of these applications, and based on different kinds of systems for data acquisition. However, pose estimation and tracking of articulated human body is related to estimating the dynamic configuration of the 3D human body (such as joints locations and joints angle) from a single image or a video sequence.

Related approaches include global and local body pose estimation, and body part estimation. Additionally, the output 3D pose information is also a rich and view-invariant representation for action recognition [Poppe, 2010]. Among the recent studies, the most extensive researches are focused on the 3D body skeleton representation, which refer to the use of the 3D joints position captured by RGBD sensor as initial input to the 3D body pose representation. Among 3D human action recognition methods, most of methods have the strategy of using a wide range of depth, pose, or image features, and do not required predefined body model. However, the approaches usually use the 3D skeleton joint configuration, depth with joints position, and multiple cues in order to represent the 3D human actions.

The extensive goal is to be able to achieve efficient action recognition applicable for, e.g., advanced human computer interaction, video surveillance, sport motion analysis automatic activity analysis and behavior understanding. We contribute to this field by providing algorithms constituting a pipeline from 3D skeleton body representations to action recognition using 3D data. The algorithms, demonstrated here, are in line with the recent development of RGBD imaging and RGBD perception.

1.1.4 Human action recognition

In computer vision, human action recognition consists to encode an action component from a video or image sequences. Action recognition has been an important topic in domains of applications include surveillance systems, video analysis, mixed reality, face analysis, object tracking and video-indexing. The problem of understanding human actions is complicated by many issues, including the fact that actions are dynamic and may not typically be recognized by simple attention to single moments in time [Lui and Beveridge, 2011]. Action recognition is further complicated by the variation between people and even between instances of a single person. However, , successful action recognition demands either explicit or implicit detect and analyze activities from sensors, e.g. a sequence of imaging, either captured by RGB cameras, range sensors, or a RGBD sensors.

The field of action representation and recognition is relatively old. In the last few decades, action recognition has been extensively researched while there is still not much mature for many real-life applications. Human action recognition has been extensively researched through methods focused on learning and recognizing actions from image sequences taken by a single monocular camera [Turaga et al., 2008]. The major issue with this data type come from the difficulty such as considerable loss of information during capture process of articulated human motion, which limits the performance of video-based human action recognition. Earlier work on human action recognition in video, stretching from human model and trajectory towards holistic and local representations methods.

Depending on the looking situation, various definitions are provided to define action or activity. [Lan et al., 2010], define action; to denote a simple, atomic movement performed by a single person. and activity; to refer to a more complex scenario that involves a group of people. Research on atomic action recognition from 3D have been proposed recently, especially after easy access of 3D data become available. In this thesis, we will cover atomic action recognition, daily activity recognition, gaming action captured by 3D RGBD camera.

The challenges of human action recognition can be enumerated into four major challenges. The first is low level challenges, the second challenge is view change. The third challenge is scale variance. The fourth challenge is intra-class variance and inter-class similarity of actions. These are the major types of difficulty of action recognition from traditional 2D image sequences. The introduction of 3D data subsequently solves these issues. Moreover, 3D data can alleviate low level challenge by providing the structure information of the scene. Despite the second challenge is solved partially by introducing multiple synchronized cameras for a traditional 2D acquisition, the problem is still inherited for recognition from range images, though, because the range image only take into account one side of the object in view. View invariant issue can be alleviated using single depth camera by providing the accurate depth map and skeletal joint information, and recognition algorithm based on these information. The scale variance can be easily adjusted in depth imaging, because the human body dimension can be estimated from the depth data. The fourth challenge remains a difficult issue for most recognition algorithms due to various type of data perspective.

1.2 Datasets for action and event recognition

In this section, we present four popular state-of-the-art 3D action recognition datasets captured by RGBD camera. and one big and very challenging dataset, which contain more challenging videos, with physiological and physical events of prematures daily care and behaviors monitoring using thermal video. This dataset is provided by the ANR project which was created with the help of medical scientists as a real benchmark to develop methods in the field of the prematures events diagnostic.

The presented datasets are used throughout this thesis work to evaluate the proposed approaches. We start with very challenging datasets containing thermal videos of infant profile during daily care of condition monitoring. For simplicity we denote this dataset by Pretherm. Then we further, go beyond this dataset, evaluate with relatively simple datasets containing videos of one person perform single action at a time, like the UTKinect-Action dataset and the Florence dataset. Although these datasets contain a small number of relatively simple actions, they have been widely used in recently for RGBD-based action recognition. Therefore, they allow us to compare the proposed techniques with many existing state-of-the-art methods. Although these datasets contain a small number of relatively simple actions, they have been widely used in recent years for evaluation.

Next, we extend an experiments on more realistic datasets. We demonstrate the MSR-3D Action dataset and the gaming 3D dataset, which contains a large set of videos with more challenging, and enumerates a complex activities of daily living and gaming.

1.2.1 Pretherm dataset

In this section, we demonstrate the real dataset that was used in this thesis. This non-public dataset is provided with the help of the project team: Pr. François Jouen, Pr. Michèle Molina, and Pr. Bernard Guillois (CHU of Caen). This project is supported by ANR project to assess the event indicators of premature infants in a realistic fashion. Physiological and behavioral events for condition monitoring assessment in premature infants are included. The dataset was captured by using a thermal camera during daily care monitoring and is composed of 30 RGB color and thermal videos of 30 neonates. The image resolution was 120×100 pixels. Events and distress behaviors series in neonates include increased body temperature, facial expression, heartbeat, and body movement. In this thesis, we focus on the problem of event recognition in premature infants by using thermal signatures of the face.

A video of each infant was recorded in one session and included physiological and behavior status monitoring. The video sequences contain the normal event in the beginning, the pain event, and post-pain event. The length of the event sequence varies between 200 and 1000 frames. The main challenges of the Pretherm dataset are: head movement of the infant, occlusion, and opening and closing the incubator. Among 30 video, we have selected 20 video from this dataset. Sample video frames from the Pretherm dataset are presented in Figure 1.1 . Figure 1.2 illustrates a series of raw thermal plots along the corresponding event to which the premature infant responded. These plots are obtained along the temporal evolution of the local facial temperature.

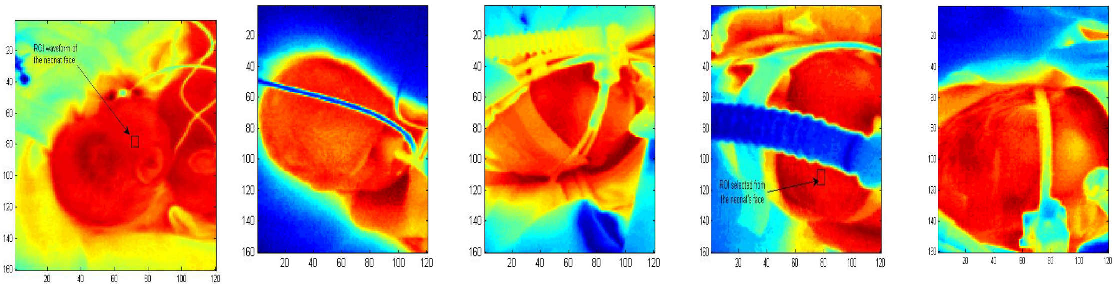

Figure 1.1: Sample Images from Pretherm dataset of the 5 infants

1.2.2 UTKinect action dataset

The UTKinect Action dataset (for simplicity the UTKinect dataset) has been introduced by [Xia et al., 2012]. It contains videos of 10 types of human actions. The full list of actions is: walk, sit down, stand up, pick up, carry, throw, push, pull, wave and clap hands. Each action is performed 2 times by 10 different subjects. The sequences are recorder using a single stationary Kinect camera with 30 frames per second.The resolution of the depth map is 320×240 and resolution of the RGB image is $640 \times 480 pixel$. In total, the dataset contains 200 action samples and the 3D locations of 20 joints were included in the dataset. The length of sample actions ranges from 5 to 120 frames. Sample RGBD images from the dataset are shown in Figure 1.3. Note that we only use the information from the skeleton joints for action recognition in our algorithm. This dataset is challenging to apply as most of the actions involve view point and high intra-class variations.

1.2.3 Florence action-3D dataset

Florence3D-Action has been introduced by [Seidenari et al., 2013]. This dataset is constructed with a RGBD sensor and includes 9 activities performed by 10 different subjects. Each subject performs every action two or three times. The full list of actions is: wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch, bow. As was suggested in [Seidenari et al., 2013] 215 sequences are used in total. The 3D locations of 15 joints are included in the dataset as well. The main challenges of the Florence dataset are: high intra-class variations (same action is performed using left hand in some sequences and

Table 1.1: The three action subsets (AS) of MSR Action 3D dataset as defined in [10]

As1	As2	As3
Horizontal arm wave	High arm wave	High throw
Hand catch	Hammer	Forward kick
Draw x	Forward punch	Side kick
Draw tick	High throw	Jogging
Draw circle	Hand clap	Tennis swing
Two hand wave	Bend	Tennis serve
Forward kick	Tennis serve	Golf swing
Pickup & throw	Side boxing	Pickup & throw

right hand in some other) and the presence of actions like drink from a bottle and answer phone which are quite similar to each other.

1.2.4 MSR action-3D dataset

The MSR Action 3D dataset has been introduced by [Wanqing et al., 2010]. This dataset contains 20 different actions, performed by ten different subjects and with up to three repetitions making a total of 567 sequences. The 3D locations of 20 joints were included in the dataset. The dataset is captured using a Kinect device. Three channels are recorded: depth maps, skeleton joint positions, and RGB video. The full list of actions is: 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 and throw. the authors was further divided the dataset into AS1, AS2 and AS3 subsets, each consisting of 8 actions as shown in Table 1.1. This was due to the high computational cost of dealing with the complete dataset. The AS1 and AS2 subsets were intended to group actions with similar movement, while AS3 was intended to group complex actions together. Most of the current methods working with this dataset have also used the subsets protocol. The reason of dividing the dataset into subsets, is that due to the high computational cost of dealing with the complete dataset.

1.2.5 Gaming Action-3D Dataset

The A Gaming Action Dataset (in short G3D dataset) has been introduced by [Bloom et al., 2012]. for real-time action recognition in gaming containing synchronized video, depth and skeleton data. Each skeleton contains the player's position and pose. The pose comprises of 20 joints. It contains of 20 types of gaming actions performed by 10 subjects, Each subject performs every action three times. The list of actions is: punch right, punch left, kick right, kick left, defend, golf swing, tennis swing forehand, tennis swing backhand, tennis serve, throw bowling ball, aim and fire gun, walk, run, jump, climb, crouch, steer a car, wave, flap and clap. In total there are 234 sequences.

1.3 Contributions and organization of manuscript

1.3.1 Problem statement and objectives

Our goal is to recognize event and 3D human actions recognition, from thermal and visual video sequences. Although a lot of methods have been proposed for event and action recognition, few results are available in the literature about thermal and 3D imaging. Similarly, it should be noted that very little works in literature has been done for multiple events recognition. This is due to the fact that event recognition, as well as activity recognition problems face a lot of common challenges, which can be described as follows:

Poor information. As opposed to traditional images, skeleton information are very poor on texture information , and it is therefore difficult to extract local features of the target object from the scene. This makes it necessary to resort to methods taking into account a several difficulties such as, lack of precision and occlusions, caused by body parts or other objects present in the scene. Therefore, in order to provide robustness to human action recognition, possible errors in skeletal data should be considered, either improving the skeleton joint representation, or relying on fusion multiple features as 3D (volume) or 2D (silhouette) depth-based information.

In contrast to 3D action, event description in thermal video such as vital signal, physiological behavior of the patients, have the limitation of foreground/background suppression, noise, and facial or body characteristic.

Occlusion. Occlusion problems are inevitable in action recognition. In a video sequence, the occlusion problem usually appears in the video areas captured from crowded environments such as airports or supermarkets. There is another kind of occlusions in activity recognition, called the self-occlusion problem, which is due to motion overlapping in the same region. In general, it is difficult to deal with occlusion problems in video processing.

Inter-class and intra-class variations. Inter-class and intra-class variations are the central issues of any pattern recognition problem. They characterize the difficulty of the problem. In action recognition, intra-class variations are often larger than inter-class variations. In the case of activity recognition, the intra-class variation between postures of the same class is usually high for many actions, e.g. a walking person may be from left to the right, from right to the left or directly facing the camera. Also, different people perform different actions in different ways, e.g. walking actions can differ in speed and stride length. In addition, the inter-class variation is low for many actions of different classes, e.g. slow running resembles jogging although the application might require to differentiate of these two activities. A successful recognition approach should be able to deal with these difficulties.

Practically, event and action recognition often consist of three main stages:

- Feature extraction and Representation: this stage first extracts features from actions of interest, which can be either global or local. Then, action is represented by the features.
- Learning models: machine learning algorithms, either generative or discriminative, are used to learn action models. The following learning methods are often employed: Support vector machines (SVM), Baysian networks, adaboost, Hidden Markov Models (HMM), conditional random fields (CRF), Extreme Learning Machine etc.
- Recognition: this step decides if action instances are present in the scene (image or video) using the models learned from the above step.

In this thesis, our main objective is to develop a new solutions to event-based thermal video and 3D skeleton-based human action recognition problems, respectively.

1.3.2 Main contributions

The goal of this thesis is the recognition of events and human actions from thermal and 3D RGBD sensors respectively.
The first part of our work is based on non redundant local features and topological persistence (TP), which are employed for events descriptions from thermal video.

According to our knowledge, there was few literature addressing the problem of events recognition of premature infant from the thermal imaging in the past. The second part of this work aim to address skeleton-based action recognition by introducing a new method for 3D skeleton joint-based human actions recognition in RGBD videos. As discussed so far, we formulate algorithms of body skeleton representation for action recognition. I summarized the contribution into the following 4 aspects:

Events recognition from thermal imaging: We propose an approaches on event recognition from thermal imaging that are able to represent the temporal evolution of thermal signature through time series, which extracted from the facial region of interest. Our approaches are based on a Temporal Local Binary Features and one dimension topological computation techniques to encode the temporal variation of thermal signature. These approaches enable the recognizer to discriminate the event in thermal video, which provide initialization of a variety of realistic tasks such as medical assessment, and preliminary health diagnosis of prematures infant in a condition monitoring system. These works were published in parts in [All alwani et. al.2014c], and [Al Alwani et.al. 2015d].

Relative angles features for pose-based action recognition: We developed a novel features for action recognition using skeletal joints information. The proposed method introduced a view invariant feature by computing the relative angles between joints in orthogonal planes. We investigate the performance of the proposed on a 3D skeleton-based action datasets. This contribution was published in [Al Alwani et al., 2014c].

Spherical harmonics for skeleton-based human action recognition: We proposed a novel temporal representation of local body joints for 3D skeleton sequence that especially maps the spherical orientations of the local joints in spherical harmonics domain and gives robust and discerning descriptions of the skeleton joints displacement over the sequence time. This spherical harmonics offer the possibility to provide the abstraction of local joint movement, which is reliable in many real-world. This contribution was published in [Al Alwani et.al. 2015a].

Spatio-temporal representation of 3-D skeleton joints-based action recognition using modified spherical harmonics: We extended the idea of spherical harmonics with spatiotemporal distribution of skeleton joints. To achieve this, we develop the real part of the SHs and use the modified part of SHs to extract the harmonics components of the skeleton joints in spatiotemporal domain. We show how to incorporate the modified SHs in spatiotemporal constraint to improve accuracy for human action recognition. This work was submitted in [Al Alwani et.al. 2015b][Al Alwani et.al. 2015c].

1.3.3 Thesis roadmap

This thesis consists of seven chapters with the current chapter being the first. The remainder of this thesis is organized in six chapters each consists of previously published work. A brief description of the consecutive chapters is as follows:
In chapter two, we review existing literature focusing on the most related and prominent state-of-the-art research techniques related to our work. In chapter three we present a new approaches for event recognition in thermal video. The new approaches are based on temporal evolution representations of thermal features using local Temporal pattern and topological persistence. We apply these approaches to encode the event in premature infant condition monitoring. This contribution can easily be applicable in diagnostic and assessment routines to help with the daily care monitoring system.

Chapters 4, 5 , and 6, present our contributions to skeleton-based human action recognition, respectively.
In chapter 4, we describe the view invariant method of skeleton joints representation using relative angles between primitive joints. In chapter 5 we give details of computing spherical harmonics on spherical orientation of joints. The proposed representation combine covariance matrix on spherical components in order to encodes the temporal evolution and dependency between local joint in explicit model. In chapter 6, we extend the spherical harmonics calculation into spatiotemporal domain for explicit model of skeleton joints. Which in a natural way encode local and global skeleton joints movements in a video sequence. In this chapter we introduce a modified SHs technique for body skeleton representation.

1.3.4 Publications

The publications that result from the work of this Ph.D. thesis are listed below.

International journals

1 **[Al Alwani et.al 2014a]** Youssef Chahir, Abderraouf Bouziane, Messaoud Mostefai, Adnan Al Alwani. Weakly supervised learning from SIFT key-points: An approach combining fast eigendecompostion, regularization and diffusion on graphs. 2014, *Journal of Electronic Imaging, Vol. 23(1)* , [ISSN : 1017-9909].

International conferences

2 **[Al Alwani et. al.2014b]** Adnan Al Alwani , Youssef Chahir. Neonatal Events Recognition using LBP descriptor and Wavelet Thresholding Technique., 2014,ICMCS14,pp 427-432.

3 **[Al Alwani et.al.2014c]** Adnan Al Alwani, Youssef Chahir, Djamal E. Goumidi, Michèle Molina, Francois Jouen. D-Posture Recognition using Joint Angle Representation. 2014, IPMU1, pp 106-115.

4 **[Al Alwani et.al. 2014d]** Adnan Al Alwani, Youssef Chahir, Francois Jouen, Thermal Signature Using Non-Redundant Temporal Local Binary-based features. Proc, ICIAR, 2014,pp 151-158.

5 **[Al Alwani et al. 2015a]** Adnan Al Alwani and Youssef Chahir. 3-D Skeleton Joints-Based Action Recognition using Covariance Descriptors on Discrete Spherical Harmonics Transform. International Conference Image Processing, 2015.

Articles under review in the international journals

6 **[Al Alwani et. al. 2015b]** Adnan Al Alwani and Youssef Chahir. Scalar Spherical Harmonics with an Extreme Learning Machine for 3D Pose-based Human Action Recognition. International Journal of Image and Vision Computing IMAVIS, 2015.(under 2nd revision)

7 **[Al Alwani et. al.2015c]** Adnan Al Alwani and Youssef Chahir. Spatio-Temporal Representation of 3-D Skeleton Joints-Based Action Recognition using Modified Spherical Harmonics. Submitted to International Journal of pattern Recognition Letters. (under revision).

8 **[Al Alwani et. al. 2015d]**Adnan Al Alwani , Youssef Chahir and François Jouen. Event Recognition in Thermal Video by Representing Temporal Evolution as Topological Persistence in Topological Space. Submitted to International Journal of Pattern Recognition. (under Revision).

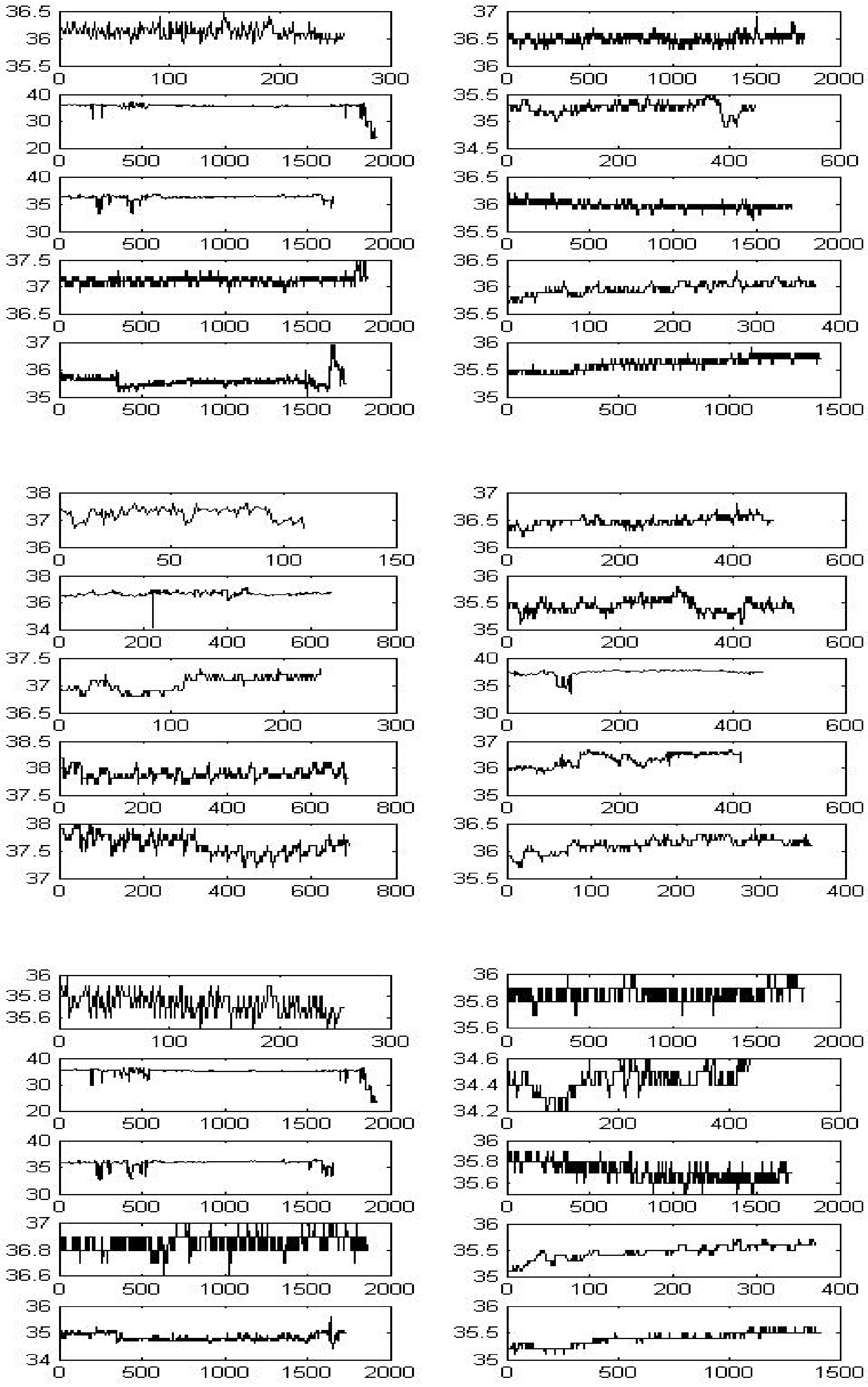

Figure 1.2: Samples of temporal evolution for10 infants responses, each signal is captured from underlying thermal signature of local facial area.(Left panel) Normal response, (Midd panel) pain response, (Right panel) Post-Pain response.

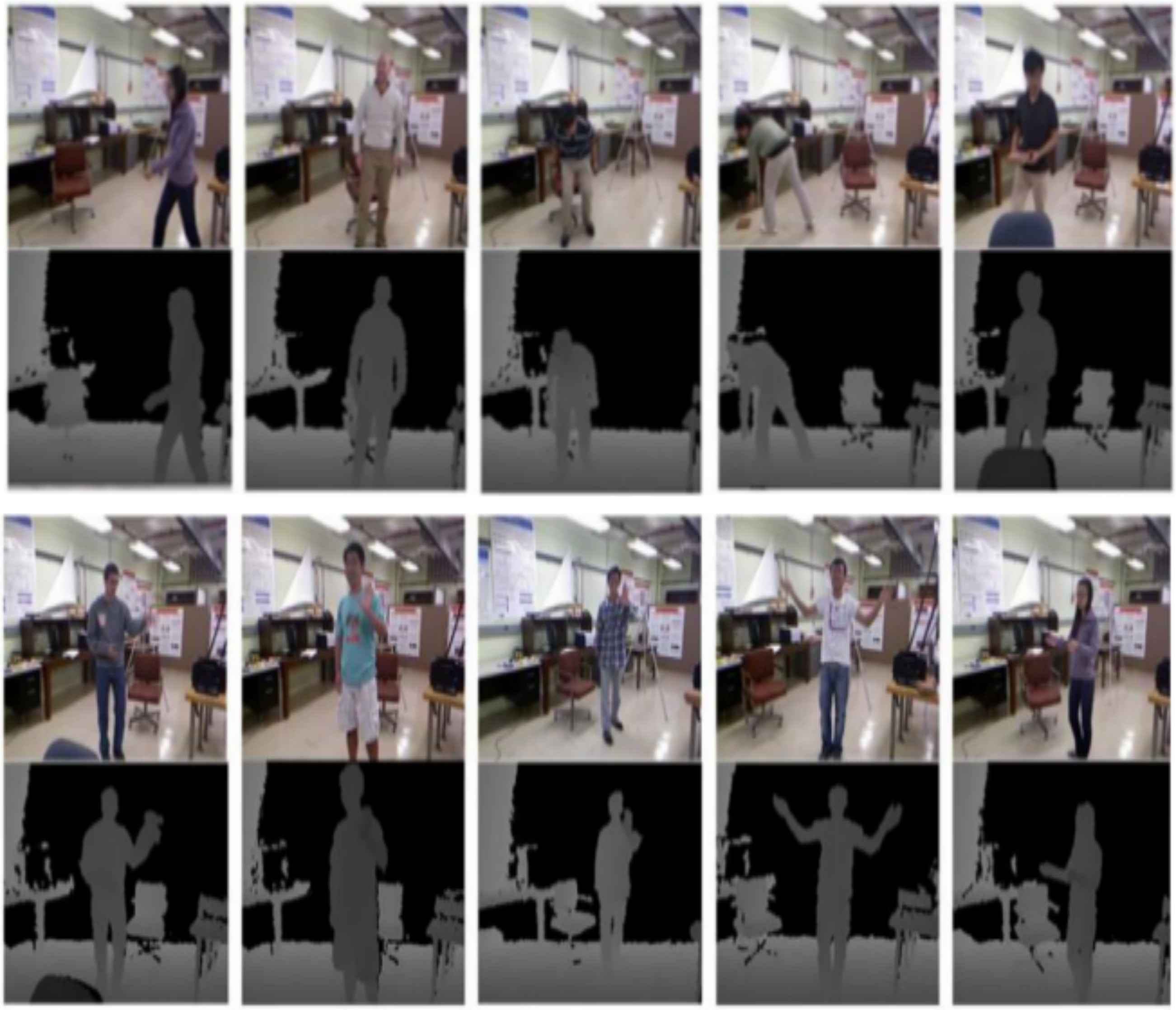

Figure 1.3: Sample images from videos of the 10 activities. RGB image frames as well as the corresponding depth maps [Xia et al., 2012]

Chapter 2

Literature review

Contents

Significant efforts of automatic event recognition has been demonstrated in different scenarios. Based on the diversity of the application areas, researchers have explored on different aspects of the problem. However, according to the demand an approaches are vary significantly. Event and Activity recognition has been studied for a long history. Moreover, Past research has mainly focused on event and activity recognition from video sequences taken by a traditional 2D single camera. a Surveys and reviews on generic action and activity recognition have been published . [Turaga et al., 2008], [Aggarwal and Ryoo, 2011], [Poppe, 2010], [Ke et al., 2013], [Chaquet et al., 2013], [Jiang et al., 2013].

In general event and activity recognition components often includes: the target domain, and environment, feature extraction and representation, and the classification task. Figure 2.1 shows a general block diagram. In contrast to visible imaging, thermal imaging is used to solve visual-based limitations. Such as poor performance with illumination variations, low lighting, poses, aging, disguise, and neuroscientists and psychologists. Numerous applications relate mainly to the particular fields of security. Such as identification, [Yoshitomi et al., 1997], object detection and recognition [Andreone et al., 2002, Davis and Keck, 2005, Dai et al., 2005, Li et al., 2012], medical diagnostic [Pavlidis et al., 2000], and health assessment [Murthy and Pavlidis, 2005] have been published. This chapter discusses the current state-of-the-arts in a range of topics. Initially, general techniques for problem solved in thermal video are examined, with discussions of interest approaches in specific domain. In addition, existing work on 3D human action recognition from RGBD is discussed, for silhouette, skeletal joint and body part locations, and local spatiotemporal features, respectively.

2.1 Related work in thermal imaging

Existing studies on feature extraction from thermal videos can be divided into two groups based on representation. The first group is structural rep-

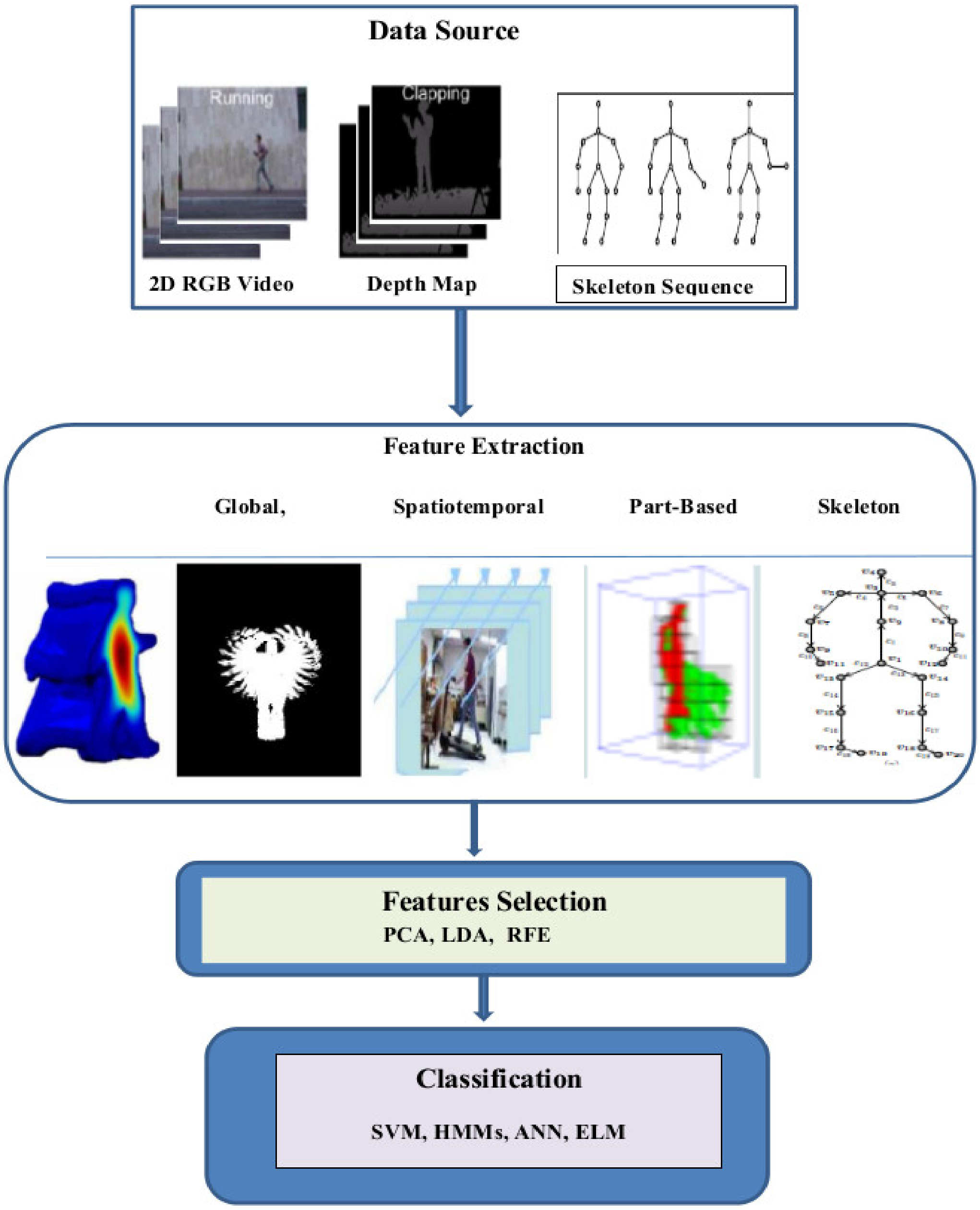

Figure 2.1: Block diagram of a typical action recognition system

resentation, where spatial holistic features are obtained using normal detection algorithms and are used in computer vision applications, such as object recognition in dark area and video surveillance. Many researchers pay their attentions to the problems of robust pedestrian detection and tracking in infrared imagery [Li et al., 2012], [Yasuno et al., 2004],[Grazia et al., 2005], [Xu et al., 2005], [Li and Gong, 2010], [Teutsch et al., 2014]. The second group is based on local and functional signals that are quantified in a temporal fashion and are used in specific domain applications. Such as health assessment, medical diagnosis, biomedical domain, and temperature monitoring system of a body. In this section, we generally review various methods with focus on feature extraction from a thermal video. The authors in [Mark et al., 2005] used three stages of pedestrian detection algorithm for a drive assistance system. In this work, the first step is identifying worm region, which removes all false positives using candidate filtering. The object in the region is validated by realizing the morphological features of an object. Contour saliency map was used in [Wang et al., 2010] to realize human detection in thermal images. Then, a template is produced from the edge samples as an improved feature vector. In [Bertozzi et al., 2003], shape context descriptor was proposed for pedestrian detection from a thermal video. Local features were used in [Jungling and Arens, 2009] to build the shape model of pedestrian detection on thermal data. A method for face identification has been developed in [Yoshitomi et al., 1997]. The method is based on 2-dimensional detection of the temperature distribution of the face, using infrared rays. The measured temperature distribution and the locally averaged temperature are separately used as input data for a neural network. While the values of shape factors are used for supervised classification

Various medical-based approaches in the field of thermal signature examination have been proposed in the literature. In [Farah et al., 2011]and [Murthy and Pavlidis, 2005], the authors studied respiration behavior-based vital signal by examining temperature changes around the nasal regions. These developments enable effective access to a significant effect in biomedical applications. Adopting face recognition techniques in medical diagnostics is a novel application area. In [Gunaratne and Sato, 2003], the author used a mesh-based approach to identify asymmetries in facial expression to determine the presence of facial motion for patients. The author in [Dai et al., 2001] proposed a method for monitoring facial expressions of patients. Temperature analysis of the face was adopted by [**?**] to explore patterns of facial stress from a distance using thermal imaging. Research in [Nhan and Chau, 2010] has recently shown a direct relation between an individual's emotional state and facial skin temperature. These variations can be reliably detected using thermal imaging.

In this thesis, we consider event detection in Condition Monitoring in Neonatal Intensive Care (NIC) settings. Condition monitoring in NIC is more challenging because of the possibilities of various problems. Simplistic assumptions in NIC event detection may no longer be valid in traditional activity settings where different behaviors and physical responses to pain such as : Crying, difficulty sleeping, agitation, frowning, and so on.

2.2 Human action recognition in RGBD sensor

Different developments of action and activity recognition, has been early devoted for human action representations from intensity images sequence. A variety of attempting are structured into three categories:
Human model based methods which employ a full 2D model of human body parts, and action recognition is done using information about positioning movements of body parts,[Moeslund et al., 2006], [Ali et al., 2007], [Parameswaran and Chellappa, 2006], [Yilmaz and Shah,].
Holistic methods, which adopts global body configuration and dynamics to represent human actions. Comparing to others approaches, holistic representations are much simpler since they only model global motion and appearance information [Yamato et al., 1992], [Bobick and Davis, 2001], [Blank et al., 2005], [Gorelick et al., 2007], [Weinland and Boyer, 2008] [Ziming et al., 2008], [Bobick and Davis, 2001]. However, since actor performs an actions in parallel to the 2D imaging camera view, thus the silhouettes-based feature extracted from 2D images are view-dependent. Also, extracting the correct silhouettes of the actor can be difficult when there is occlusion or bad lighting conditions.
Local feature methods: local features characterize an appearance and motion information for a local region in video. Such features are usually extracted directly from video without additional motion segmentation or human detection [Laptev and Lindeberg, 2003], [Laptev,], [Harris and Stephens, 1988], [Geert et al., 2008], [Wong and Cipolla, 2007], [Dollár et al., 2005], . However, here we demonstrate the related works on human action recognition from RGBD images. To this end, as mentioned above. RGBD capture a depth image (D), along with an RGB image , altogether gives RGBD. Based on the features used, a depth image can be further provides a 2D silhouette, a 3D silhouettes, or a

skeleton model. In what follows, we discuss 3D silhouette and skeleton based approaches, respectively.

2.2.1 Action recognition from 3D silhouette

In a RGBD sequence, the global shape of a human can usually be identified more easily and accurately. In addition, the depth image provides both the body shape information along the silhouettes, and the whole side facing the camera. That is, depth images provides more information about body silhouette. Inspired by representations built from 3D silhouettes, many algorithms have been proposed for action recognition. [Wanqing et al., 2010] construct a bag of 3D points from contours of the projections of the 3D depth map to obtain a set of action information . In order to reduce the size of the feature vector, the method selects a specified number of points at equal distance along the contours of the projections. [Bingbing et al., 2013] extend the original MHI to a three-Dimensional Motion History Image (3D-MHI). Two additional channels of forward-DMHI and backward-DHMI are equipped which encode forward and backward motion history. [Xiaodong et al., 2012] also project depth maps onto three orthogonal planes and accumulate the whole sequence generating a depth motion map (DMM). Histograms of oriented gradients (HOG) are obtained for each DMM . [Fanello et al., 2013] propose a global Histogram of Oriented Gradient (GHOG) based on the classic HOG [Dalal et al., 2006], which was proposed for human detection from RGB images. The GHOG describes the visual appearance of the global silhouettes without splitting the image grid into cells. The highest response of the depth gradient on the boundary contours reveal the pose of the person. [Ballin et al., 2012] proposed a 3D grid-based descriptor to estimate the 3D optical flow related to the tracked people from point cloud data. Relaying on the combination of silhouette shape and optical flow in the same feature vector, A popular feature is proposed by [Tran et al., 2008]. In their work, radial histograms of the silhouette shape and the axises of the optical flow are encoded.

2.2.2 Action recognition from skeletal data

Relying on the articulated nature of the human body, the human body consisting of a set of rigid segments connected by joints,and human motion can be considered as a continuous evolution of the spatial configuration of these rigid segments [Zatsiorsky, 1998]. In computer vision, Existing skeleton-based human action recognition approaches either focused on extracting the joints or detecting body parts and tracking them in the temporal domain.

Inspired by the algorithm proposed in [Shotton et al., 2011], Shotton et al. propose to register the 3D body joint position from a depth image. Resulting an easy way to handle the skeletal joint locations for action recognition with better accuracy. The 3D skeleton joint-based approaches have been explored by various researchers. [Yao et al., 2011] indicated that the application of skeleton data (e.g., positions, velocities, and angles of a joint from a human articulated body) outperforms gray-based features captured by 2D camera in an indoor environment scenario. In general, many useful features can be initially extracted from RGB-D skeletal data. The majority of these features can be divided into two: those that are based on the angular characteristics of joints and those that are based on the generic 3D coordinate of joints.

In certain action recognition methods, the features are developed in complex models to form the representation of the motion sequences. The 3D joint positions are commonly extracted as features through four mechanisms. First, raw 3D data are recognized directly without any further processing [Raptis et al., 2008, Shimada and Taniguchi, 2008, Wang and Lee, 2009]. Second, these data are further processed to address certain challenges [Barnachon et al., 2013, Wang et al., 2012b, Zhao et al., 2013]. Third, the distances between each joint can be used as a distance-based feature vector for each frame [Antônio et al., 2012]. Fourth, the features for the selected joints can be simply calculated with reference to the relative distance between joints [Wang et al., 2012b].

In [Hussein et al., 2013], the human body skeleton was interpreted by directly constructing 3D skeleton joint locations as a covariance descriptor, and the temporal evolutions of the action dynamic were modeled using a temporal hierarchy of covariance descriptors. In [Lv and Nevatia, 2006a], the 3D coordinates of the joints were used for a skeleton representation of the human body. Correspondingly, the temporal nature of the action sequence was modeled with a generative discrete hidden Markov model (HMM), and action recognition was performed using the multi-class AdaBoost. The view-invariant representation of the human skeleton was proposed in [Xia et al., 2012] by partitioning the 3D spherical coordinates into angular spaced bins based on the aligned orientations with respect to a coordinate system registered at the hip center. A generative HMM classifier, which addresses the temporal nature of pose observations, was then used to classify each visual code word identified with the cluster method. The proposed work in [Wang et al., 2012b], applied the idea of the pairwise relative locations of joints to represent the human skeleton. The temporal displacement of this representation was characterized using the coefficients of a Fourier pyramid hierarchy. Moreover, the researchers proposed an action let -based approach, in which the effective joint combinations were selected using a multiple kernel learning approach. In [Yang and Tian, 2012], the skeleton joints were represented by combining the temporal and spatial joint relations. To explicitly model the motion displacement, the researchers adopted a method for skeleton representation on the basis of relative joint positions, temporal motion of joints, and offset of joints with respect to the reference frame. The resulting descriptors were projected onto eigenvectors using PCA. In this case, each frame was described by an EigenJoint descriptor, and action recognition was performed using the nave Bayes nearest neighbor. The same scheme was used for the skeleton representation in [Zhu et al., 2013], in which action recognition was achieved by adopting the random forest classifier. The view-invariant action representation framework was proposed by [Evangelidis et al., 2014]. In this work, the skeletal quad-based skeletal feature was adopted to encode the local relation between joints in quadruple form. Consequently, the 3D similarity invariance was achieved. The researchers also adopted a Fisher kernel representation based on a Gaussian mixture model. Such a representation generates the skeletal quads and invokes a multi-level splitting of sequences into segments to integrate the order of sub-actions into the vector representation. In [Vemulapalli et al., 2014], a human skeleton was presented as points in the Lie group. The proposed representation explicitly models the 3D geometric relationships among various body parts using rotations and translations. Given that the Lie group was a curved manifold, the researchers mapped all action curves from the Lie group to its Lie algebra, and the temporal evolutions were modeled with DTW.

Angular direction of joint can be computed, which is invariant to human body size and view. The work proposed by [Sempena et al., 2011], adopt joint orientation along action sequence to build a feature vector and apply dynamic time warping onto the feature vector for action recognition. [Bloom et al., 2012] concatenates a variety of features like: pairwise joint position difference, joint velocity, velocity magnitude, joint angle velocity magnitude, and joint angle velocity. Altogether, 170 dimensional features vector was formed in order to recognize gaming action. In [Chaudhry et al., 2013], a human skeleton was hierarchically grouped into local parts and each part was characterized using bio-inspired shape features. The temporal nature of these features were encoded using linear dynamical systems. Pairwise affinities between joint angle trajectories was introduced in [Ohn Bar and Trivedi, 2013], for skeletal sequences representation. A sequence of the most informative joints (SMIJ) per action instance was presented by [Ofli et al., 2014]. This selection is based on joint related measures such as the moments of the joint angles. Several encoding methods are suggested for the vector representation of SMIJ.

Researchers also tried the local occupancy pattern to treat an action sequence as a 4D shape. In this 4D space the 3D orthonormal space around the joint is partitioned into spatial cell. The number of points that fall into each cell are counted to obtain the occupancy feature of that cell in certain time range. [Wang et al., 2012a] defined the random occupancy patterns in the $(x; y; z; t)$ domains, their work combine the skeleton joints features and local occupancy features to recognize activities. Partitioning the whole 4D space into sub-volume was proposed by [Vieira et al., 2012], the authors extracted occupancy patterns from every partition. hence, the local occupancy pattern is quite sparse. Thus, a modified-PCA called Orthogonal Class Learning (OCL) is employed to reduce the length of the feature.

Part I

Event recognition in Neonatal Intensive Care (NIC) from thermal imaging

Chapter 3

Event recognition in Neonatal Intensive Care system

Contents

The event recognition method proposed in this chapter is based on temporal feature descriptors that are locally extracted from a thermal video. As illustrated in Figure 3.1, our proposed method includes four main steps:

1) Preprocessing and raw feature extraction step that rotates the images, resizes the images to 100×120 pixels, and crops the images to define only the facial area. Patches of size m x m pixels are defined over the facial area, and multiple channels of temporal evolution of the temperature values are calculated from these patches.

2) A non-redundant-base local feature selection step that calculates the local temporal evolution descriptors from a set of selected channels.

3) Feature selection based on topological persistence (TP) step that calculates the critical values of a 1D signal.

4) Event recognition step where feature descriptors are used to discriminate the events by means of extreme learning machine (ELM) and SVM classifiers.

3.1 Spatio-temporal dynamic texture descriptors

3.1.1 Preprocessing and raw features channels

In video preprocessing, the original facial images are first tracked and extracted. After locating the subject image, the images are thresholded to identify the image foreground from the background. This procedure is achieved by assuming that the facial temperature distribution is relatively higher than the background. The images are cropped to include only the facial region (Figure 3.1) to facilitate the computational procedure of the approach. Then, patches with a dimension of 20×20 pixels are extracted from the tracked facial region. The preprocessing

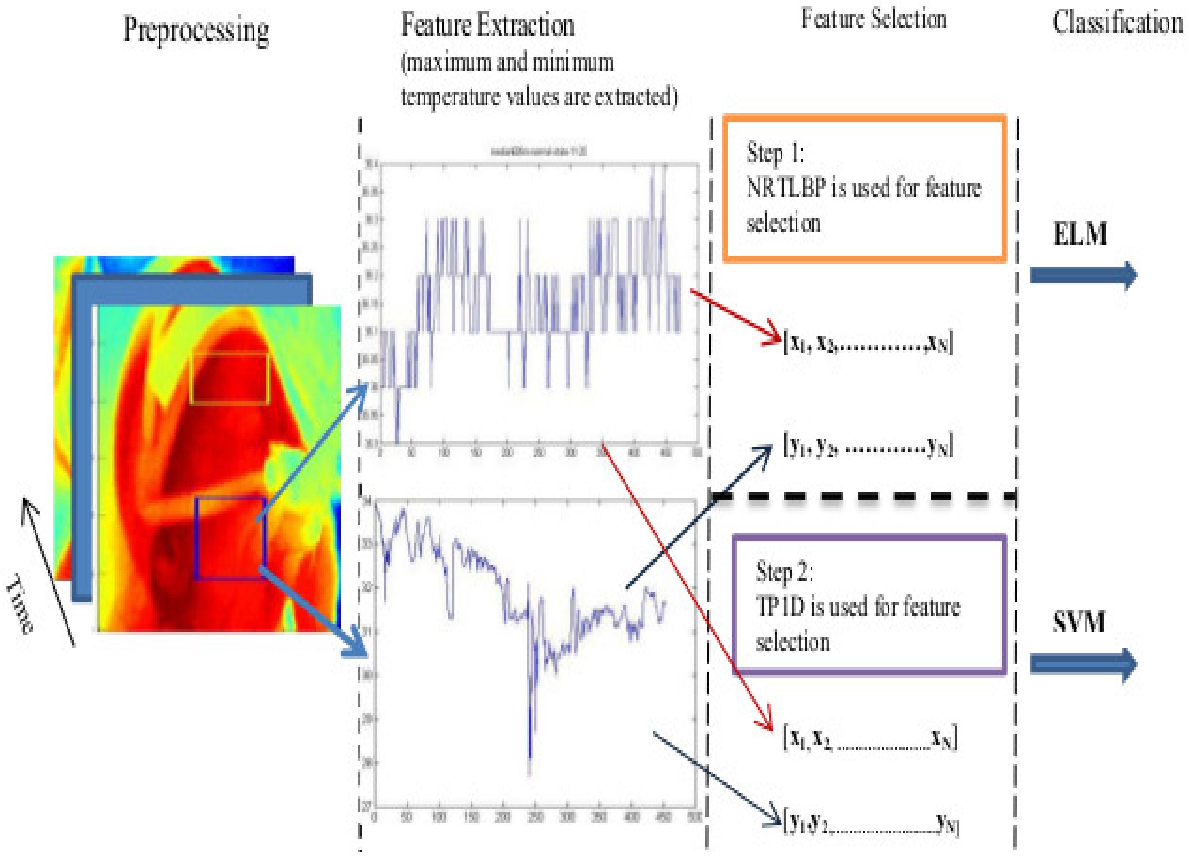
Preprocessing
Feature Extraction
(maximum and minimum
temperature values are extracted)
Feature Selection
Classification
Time
Step 1:
NRTLBP is used for feature
selection
ELM
$[x_1, x_2, \ldots\ldots\ldots\ldots x_N]$
$[y_1, y_2 \ldots\ldots\ldots\ldots y_N]$
Step 2:
TP ID is used for feature
selection
SVM
$[x_1, x_2 \ldots\ldots\ldots\ldots x_N]$
$[y_1, y_2 \ldots\ldots\ldots\ldots y_N]$

Figure 3.1: Experimental design

step emphasizes the dependency of the event's information contained within different regions of the face, such as cheeks and foreheads. In accordance with our initial study in [Al Alwani et al., 2014], two raw channels of temperature values are temporally extracted from the patches, which are defined over the facial region. These raw features are the maximum and minimum temperature values. Furthermore, we establish two raw signals for the three-condition events, which characterize thermal signature in monitoring state. For instance, three events adopted in this study are normal, pain, and post-pain, respectively. The time series of each event is shown in Figure 3.2. The figure indicates that the response nature of each event is initially characterized by the signal temporal evolutions of that event. However, collecting multiple raw channels of facial samples will obtain the best recognition rates. The raw channels of temperature values are used as initial input features for the proposed descriptors.

3.1.2 Thermal signature using NRTLBP-based features

In this section, we propose a method for event recognition NIC system from thermal signature based 1-D signal. We use the Non-Redundant Temporal Local Binary Pattern (NRTLBP) as a descriptor of for the signal Pattern Of Interest (POI) signal. Moreover, We assume that the layout of subjects is considered for all the cases in the front view, and ROI is defined over the subject face. In addition, we build raw thermal signatures for all subjects samples. This is achieved first by defining local patch as shown in the recognition system which illustrated in the Figure 3.1. Then the maximum Max and the minimum Min values are computed from the local patches along the video sequences. For more illustration, we compute the Max and Min temperature values and denote it as raw thermal signals (we call it raw signals for abbreviation). The raw signals quantify three condition events, which characterize the monitoring state during daily care. For instance, normal event, pain event and post-pain event are three events adopted in our study. We apply the NRTLBP descriptor on the raw thermal signal in order to extract efficient feature vector. To make the descriptor robust against minor temporal variation and noise, the wavelet decomposition of the raw signal is used in order to extract the approximation wave-components. Then, NRTLBP is applied on the wave-components which further provides feature descriptor in wavelet domain (WNRTLBP). We provide an evaluation of our method using Support Vector Machine SVM and the real dataset (Pretherm dataset) composed of thermal videos developed in the context of Infant pain project (a french project supported

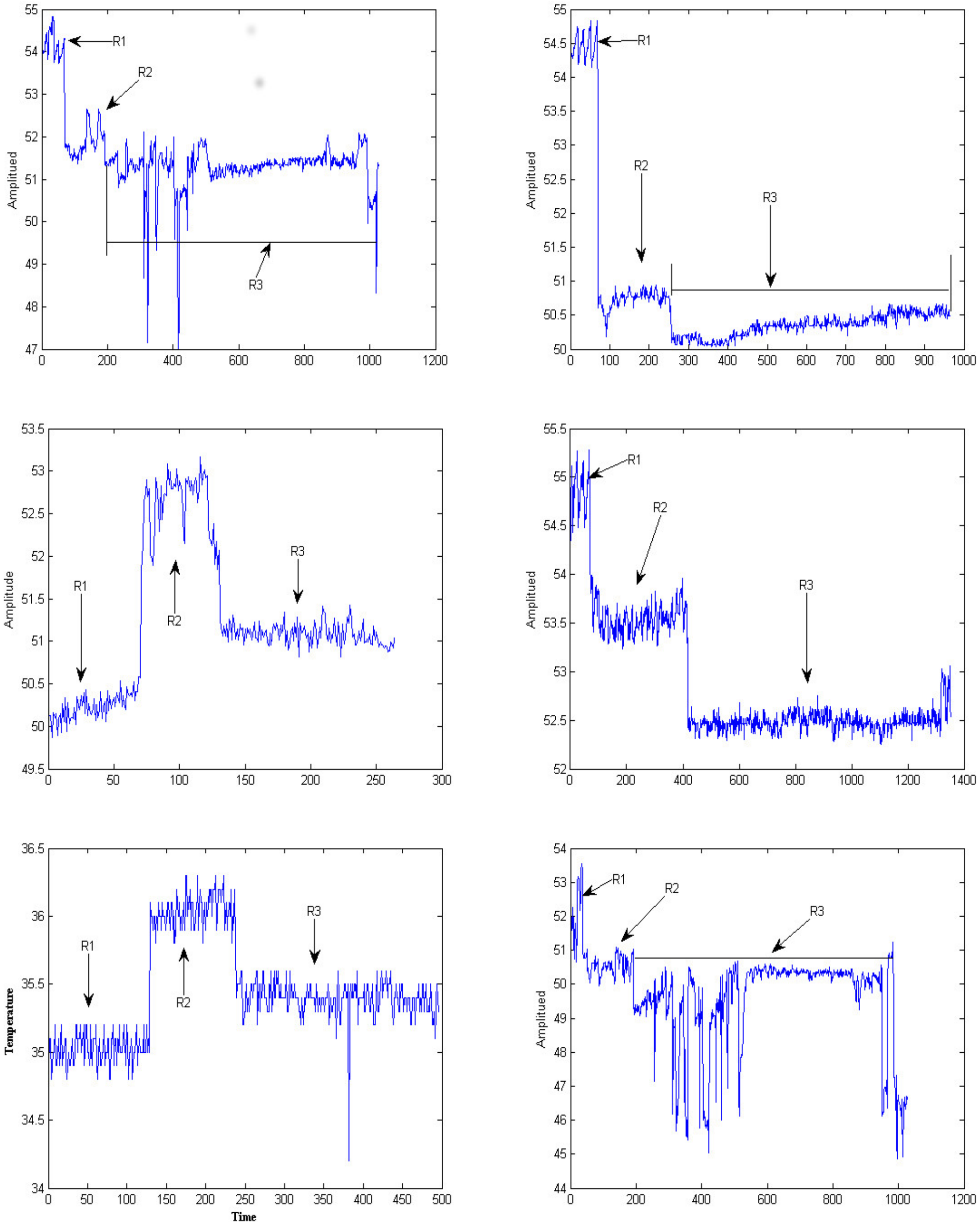

Figure 3.2: Six examples of raw features of six infants response, each panel consist of three segment. R1, indicate the normal, R2 indicates pain response, and R3 indicates post-pain response.

by the French National Research Agency Projects for science ANR). Experiments show that this algorithm achieves superior results on this challenging dataset.

Non-Redundant Temporal Local Binary Pattern-based features

Once the raw thermal features are extracted, the video sequences is considered as a collection of raw signals. We denote by $X_i = T_1, T_2, \cdots, T_n$. $i = Max$ or Min, and T is the temperature value at each patch. In order to represent the raw signal in efficient manner we propose the Temporal Local Binary Pattern descriptor, which is an extension of original 2-D LBP operator into the temporal domain.
Normal LBP has originally been proposed for texture analysis and classification [Ojala et al., 2002]. Recently, it has been applied on face recognition [Timo et al., 2004] and facial expression recognition[Ahonen et al., 2006].
The TLBP operator labels the samples of a signal by thresholding a center sample against neighborhood set within defined window. We denote by $x[n]$ the sampled signal, and w is the size of samples window. The TLBP operator on a sample $k \in w$ in is given by:

$$TLBP_w(f[k]) = \sum_{q=0}^{\frac{w}{2}-1} \left\{ sign[f[k+q-\frac{w}{2}] - f[k]]2^q + sign[f[k+q+1] - f[k]]2^{q+\frac{w}{2}} \right\} \tag{3.1}$$

where the sign function is given by:

$$\text{Sign} f[k] = \begin{cases} 1 & \text{if } f \geq Thr \\ 0 & \text{if } f < Thr \end{cases} \tag{3.2}$$

and where q is the number of sampled points (neighbor samples of k) whose the distances to k do not exceed the window w. For a M block, a TLBP histogram of 2^q bins is computed for feature representation.
Despite that the outcome of the TLBP has been successfully encoded the 1D signal-based applications. It has limitation when adopted to address the high variability of local neighborhoods samples. Also the storage ability of TLBP is considered as a main disadvantage, for instance, the TLBP actually requires the $2K$ of histogram structures for each segment. In the other hand, LBP is sensitive relatively to the slowly varying signal. Moreover, at a critical points, the difference will be relatively large, whereas at transition, the differences strictly depends on the directions of the edge transition. Another aspect of TLBP is called no-redundant NRTLBP. Intuitively, the NRTLBP considers a TLBP code and its complement as the same [Nguyen et al., 2010]. It contains information about the distribution nature over the whole signal and characterizing a statistical description of signal.

In this work, we adopt a novel extension of the TLPB, called Non-Redundant LBP, in order to address both of the aforementioned challenges. The NRLBP is defined by:

$$NRTLBP(f[k]) = min[TLBP(f[k]), 2^q - 1 - TLBP(f[k])]. \tag{3.3}$$

Having obtained the NRTLBP for the whole signal, a histogram vector of the NRTLBP code is used as a final feature descriptor. Obviously, the number of bins in the NRTLBP histogram is reduced to the half. Furthermore, compared with the original TLBP, the NRTLBP provides more discriminative power, and encodes uniformly the change between the signal samples. Hence, the NRLBP is more compact and adaptive with the dynamic shape of the signal at various instances and levels.

NRTLBP of approximated coefficients

The estimated raw feature is subject to error and significant noise. With the facial region, we may acquire the physiological response in neonatal intensive care with redundant, mostly weaker and reshaped signal, inaccurate estimations occur when parts of the facial is occluded and presence of large motion artifact. We further extend the idea of NRTLBP into a wavelet domain in order to reduce the noise level, the redundant, an perform efficient feature extraction.
The method is based on using multi-resolution wavelet transform WT to decompose the raw features and build new feature representation. However, WT doesn't use any window an tries to decompose a signal into wavelet basis. Moreover, the WT requires no prior about the samples distribution and provides low computational cost. The wavelet transform $C(j, k)$ of a finite-energy signal $f(t)$ is defined as its scalar product with the wavelet $\Psi_{jk}(t)$ [Mallat, 2008]. In other words, the wavelet transform represents the correlation of the signal $f(t)$ and the wavelet $\Psi_{jk}(t)$ as:

$$a_{jk} = \sum_{t} f(t)\Psi_{jk}(t). \tag{3.4}$$

Where $\Psi(t) = 2^j\Psi(2^jt - k)$ is the mother wavelet corresponds to scale $j = 1$, k is the translation factor, and j is the scale parameter. A wavelet representation of a function consists of, a coarse overall approximation a^j, and detail coefficients that influence the function at various scales. Therefore, the approximation coefficients are considered as a feature extraction vector in a wavelet-domain. Then using NRTLBP for feature description. The smooth change property of the approximation features is illustrated in the Figure 3.3 where each sub-figure shows the

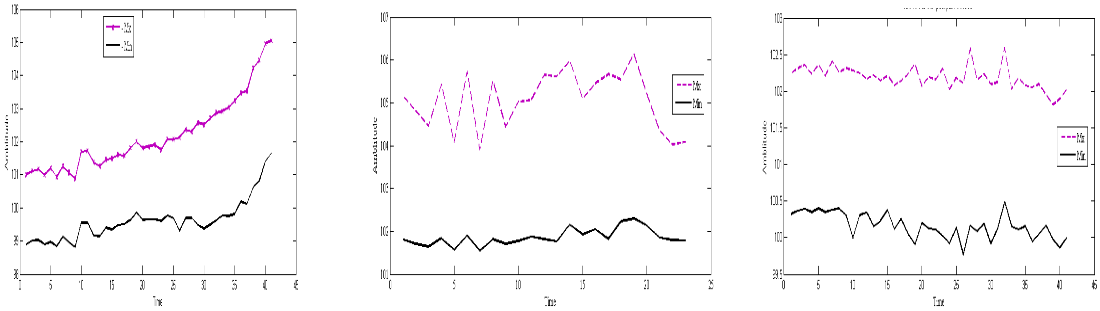

Figure 3.3: Approximation component of raw features correspond to three kinds of events. from left to right, normal event, pain, and post-pain event signatures.

feature in wavelet-domain for the corresponding event in the time domain.

Table 3.1: Recognition rate results, from the Max and Min raw features that are directly applied into SVM

Events	Recognition rate% (Max Temp.value)	Recognition rate% (Min Temp. value)
Normal Response	50	60
pain Response	90	60
Post-pain response	50	70
Overall	63.33%	63.33%

Table 3.2: Recognition rate results, using NRTLBP based raw feature descriptor

Events	Recognition rate% (Max Temp.value)	Recognition rate% (Min Temp. value)
Normal Response	80	50
pain Response	90	90
Post-pain response	80	90
Overall	83.33%	76,67%

Table 3.3: Recognition rate results, using WNRTLBP based raw feature descriptor

Events	Recognition rate% (Max Temp.value)	Recognition rate% (Min Temp. value)
Normal Response	90	100
pain Response	100	95
Post-pain response	85	93
Overall	91,667%	96%

3.1.3 Experiments

In this section, we evaluate the proposed method by experimenting with the Pretherm dataset . Due to non available of the benchmark related to our method, we did not attempt a comparison with other methods. Rather, we provide results only to show that our algorithm works well on a real data.

The premature infant data involves 20 neonates videos. Each includes the subject behaviors during daily condition monitoring (event and condition are used interchangeably). All events are performed for each subject and notated by the clinical during condition monitoring. From each video we select the set of events which includes normal health event, pain event, and post-pain event. A total of 60 clips were extracted ranging from N occurrences of normal event, to M occurrences of pain event and so on.

We divide the clips into three groups of 20 video each. For instance, each group divided into 2 subset of 10 clips each. In the training phase, we use half of the group and the normal was used for the test task.

We consider a recognition task for three class problems including the normal event, pain event and post-pain event respectively. Furthermore, two features descriptors are used in the experiments. The first one is the descriptor of raw thermal signal, we denote it by NRTLBP and the second one is the descriptor based on the approximation components, and we denote it by WNRTLBP. The number of neighboring samples was set to 4 samples for NRTLBP codes, and the Daubechies wavelet with one level decomposition was used for approximation coefficients extraction. We performed the evaluation of our method using linear SVM.

In all experiments, first subset of each group was used for training and the second subset was used for testing.

We started the experiment directly by providing the raw features (Max, and Min raw features) into classifier. The recognition results are in the form of recognition rate, the results of the first experiment are reported in the Table 3.1.

We also run the test by using NRTLBP descriptor based on two types of raw features. The results are shown in Table 3.2.

In order to be able to do a full comparison of methods, WNRTLBP based experiments are performed in the same manner as above. Results from this experiment are presented in Table 3.3.

As can be seen in Table 3.1, the overall performance of both raw features (i.e. Max and Min values) is poor and did not improve the discrimination between the events. Moreover, the direct Max and Min recognition have a lower performance than NRTLBP and WNRTLBP. Both Max and Min did not provides local dependency between signal components.

If we compare the results of using the NRTLBP and WNRTLBP descriptors, we see that the results for both of these descriptors are comparable for individuals classes. The over all performance of the WNRTLBP is 8.33 % better than NRTLBP in the case of using the Max raw feature, and 20.33 % better than NRTLBP when Min raw feature is used. As we can see, the approach based on the wavelet achieves a good accuracy rate.

It is interesting to note that even this is preliminary study of NIC pain response, the results from tested real dataset indicate that the three classes could be separated quite well from each other.

3.2 Topological-based temporal features descriptor

We employ topological computation as a novel 1D signal characterization method for event recognition in neonatal intensive care system. Topological Persistence (TP) is used to characterize temporal evolution of the 1D signal based on its critical attributes. To explore the informative events from a 1D signal, we first calculate the critical values for each data channel. Then, these critical values are paired according to the principle of topological persistence to obtain the abstract profiles of each raw signal.

However, TP 1D is a class for topological attributes that are used to find local critical points and their persistence in 1D data. Moreover, local minima and maxima are extracted, related to each other, and sorted with respect to their persistence. The local minima and maxima values of a signal approximately characterize the shape abstract of the signal by using only a set of a reduced number of critical values. By pairing these local critical values in a simple fashion, we can construct the thermal signature representation.

3.2.1 Topological persistence

Topological persistence has been investigated recently in computer graph modeling [Weinkauf et al., 2010, Brechbühler et al., 1995]. We refer the readers to

[Edelsbrunner and Harer, 2008] for an excellent review of persistence homology and to [Hatcher, 2002] for an introduction to simplicity homology.
Let $X_i = (T_1, T_2, ..., T_N)$ denote the ith raw channel of 1D thermal feature and $i = [1, \dots .k]$. Let k denote the length of the video sequence. We denote by $\zeta : R \rightarrow R$ a smooth function, with an extrema component w. If the first-order derivative of the function ζ is equal to zero, that is, $\partial\zeta(w) = 0$, then $\zeta(w)$ has a critical value of w. Each critical point is then composed of either a local minimum or a local maximum [Edelsbrunner and Harer, 2008]. We define each $s \in R$ that satisfies $\zeta(w) \leq s$ as a sub level set. We also denote by $\beta_i, \alpha_i, i = [1, \dots, m]$ the calculated local critical extrema. Then, we arrange the local maxima and minima from the smallest to the largest value as:

$$\beta(1) < \beta(2) \cdots < \beta(m), \alpha(1) < \alpha(2) \cdots < \alpha(m). \tag{3.5}$$

At a local minimum the sub level set have birth a new component i.e.

$$M(\alpha_i) = M(\alpha_i - \delta) + 1. \tag{3.6}$$

In the same sense at a local maximum, the sub-level set has the death of a component. Two components then merge into one component. i.e.,

$$M(\beta_i) = M(\beta_i - \delta) - 1. \tag{3.7}$$

Where δ, the small increase of sub level components. As an example, the critical values of the signal is illustrated in Figure3.4, it can be noted that the critical points can be used to reliably characterize the different events.

On the basis of the topological theory, the topological attributes of a topological space are abstracted by the local variation at extreme points of a smooth function on that space [Milnor, 1973]. However, the topology space produces pairs (α_i, β_j) of critical values such that a new component is generated at α_j and vanishes at β_i. The critical points calculated from the above procedure are paired by the following rule [Edelsbrunner et al., 2000]; When a new component is introduced with the local minimum, the new component is identified. At the same time, when we pass a local maximum with merged two components, we pair the maximum with the higher of the two local minima of the two components. According to the paired rule, the extreme points that are paired do not necessarily need to be contiguous. Similarly, other critical values of the function are paired in the same

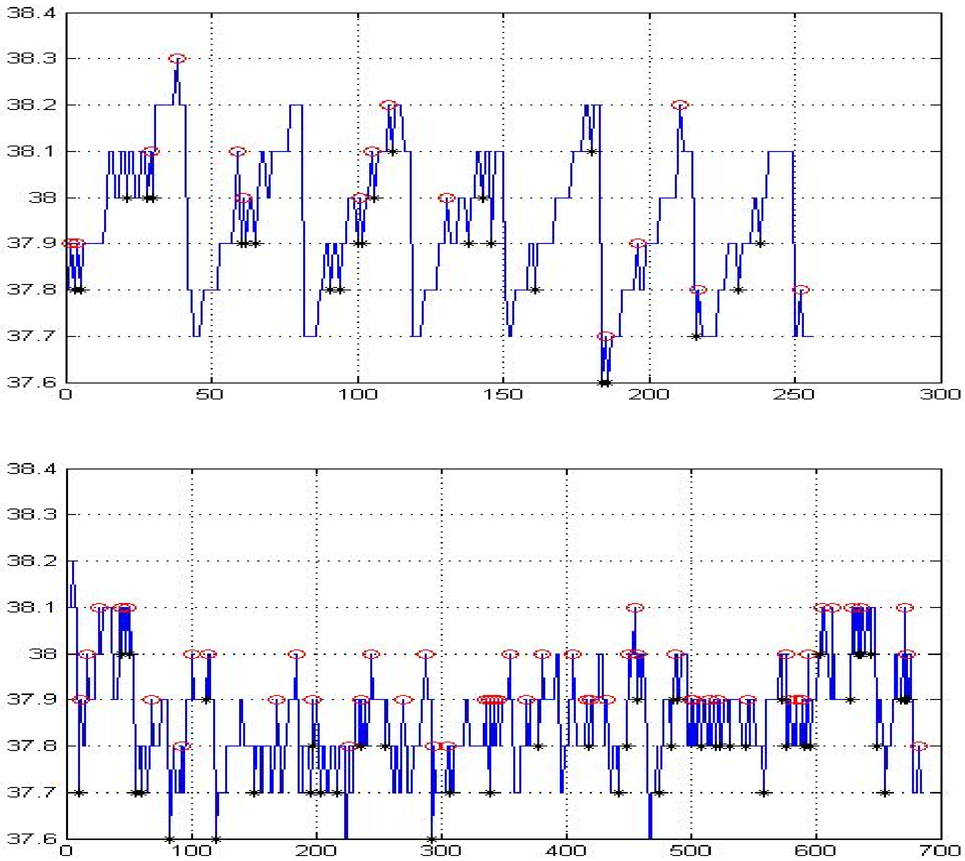

Figure 3.4: Critical points selection from raw thermal signatures which represent temporal responses of subject sample. Top, normal event, and Bottom, pain events

rule. The procedure of paired critical values is detailed in Figure3.5.
Obviously, the paired extremes vector is the topological parameter that approximately characterizes the events from a thermal signal. The computed extremes points $TP = [p_1, p_2, \cdots, p_N]$ are collected in a vector that represents the respective event feature vector. Before training or testing, TP feature descriptors are normalized to have the same features length.

3.2.2 Extreme Learning Machine

Event recognition is performed in this section by using two classifiers to assess the performance of the proposed method. ELM and linear SVM are used to classify the final feature vectors. The classification task is beyond the scope of this paper. We only introduce ELM as a new classification. Intuitively, ELM efficiently provides high-learning accuracy and faster training time compared with other learning al-

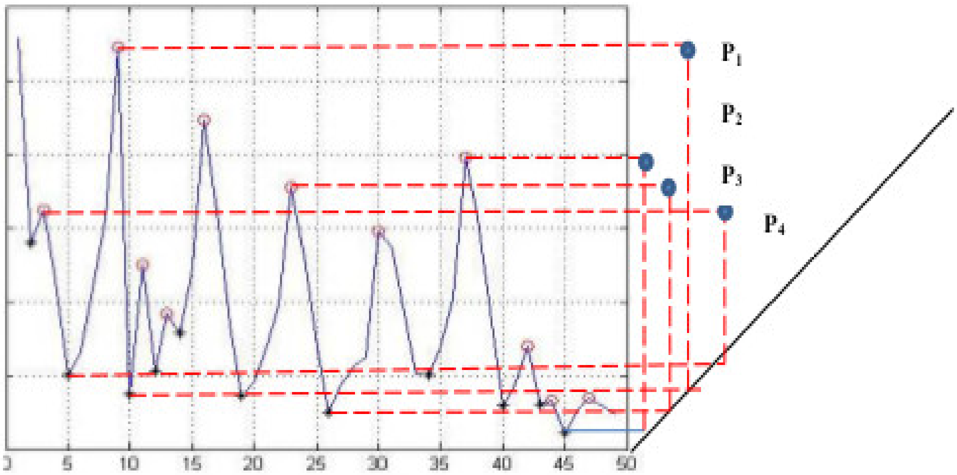

Figure 3.5: A single variable function with N local minima and local maxima. The critical points are paired and ordered to form the topological persistence

gorithms. Recently, ELM has been extensively devoted to learning single hidden layer feedforward neural networks (SLFNs) [Huang et al., 2006]. Hidden nodes in ELM are randomly initialized and do not have to be iteratively tuned. Essentially, the hidden nodes in ELM remain fixed after initialization, while only the input weight parameters need to be learned. ELM was successfully adopted by [Minhas et al., 2010] for human activities recognition from video data.

Let $x_i, y_{i}{}_{i=1}^{Q}$ denotes the training samples where $x_i \in R^N$ and $y_i \in R_M$, the generalized SLFN output of ELM is

$$y_i = \sum_{j=1}^{L} \omega_i \psi_i(\tau_j \cdot \chi_i + \lambda_j) = \mathbf{\Omega\Psi}, \quad Y \in R^Q * M. \tag{3.8}$$

where $\mathbf{\Omega_j} \in \mathbf{R^{L*M}}$ is the output weight vector parameter, $\psi(.)$ is the activation function, and $\tau_j \in R^d, \lambda_j j \in R$ are the hidden node parameters. According to the learning rule of ELM [Huang et al., 2012], both hidden node parameters are randomly assigned during the learning phase. During the linear parameter solving stage of ELM, the weight parameters denoted by ω are solved by minimizing the training error sense as follows:

$$min\|\mathbf{\Omega\Psi} - \mathbf{H}\|, \Omega \in R^{L*M}. \tag{3.9}$$

Where $\boldsymbol{\Psi}$ is the hidden layer output matrix:

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1(\chi_1) & \cdots & \psi_L(\chi_1) \\ \vdots & \vdots & \vdots \\ \psi_1(\chi_N) & \cdots & \psi_L(\chi_N) \end{bmatrix}. \tag{3.10}$$

and $\mathbf{H}$ is the training data matrix denoted as

$$\mathbf{H} = \begin{bmatrix} h_{11} & \cdots & h_{1m} \\ \vdots & \vdots & \vdots \\ h_{N1} & \cdots & h_{Nm} \end{bmatrix}. \tag{3.11}$$

Assuming that the number of hidden neurons L is less than the number of the training set (i.e., $L < Q$), the optimal solution to minimize the training error is given [Huang et al., 2012] by

$$\boldsymbol{\Omega}^* = \boldsymbol{\Psi}^*\mathbf{H}. \tag{3.12}$$

Where $\boldsymbol{\Psi}^*$ is the Moore-Penrose generalized inverse of matrix $\boldsymbol{\Psi}$ [Huang et al., 2012].

In the training phase, the input feature vectors that belong to a set of actions are expressed in terms of data matrix. Each row of the data matrix represents a specific action, and the corresponding column vector represents the feature vector.

3.2.3 Experimental design

This section describes the experimental results using the proposed feature descriptions and two classifiers (ELM and SVM) to classify the 60 video clips in Pretherm dataset into three classes, namely, normal, pain and post-pain. We also provided results to show that our algorithm works well on real data. We tested the performance of the proposed method on 60 thermal clips captured from 20 subjects. These clips were divided into three classes: 20 clips of normal, 20 clips of disease, and 20 clips of post-pain responses. A total of 20 subjects from the Pretherm dataset were selected. The minimum number of frames that are available for each clip ranged from 100 to 400 frames.

Actual normal, pain, and post-pain modes for each infant were used to train ELM and SVM classifiers. For each subject, two local regions that correspond to the face and the front area of the facial were selected. In the selection process, two raw channels were selected, where the infant facial expression reflected a normal condition, distressed intensity for pain, and a calm response for post-pain. Recognition tasks of three classes problems were considered. In the experiments, we labeled the extracted feature vectors as NRTLBP and TP1D descriptors. In the classification stage, ELM and SVM experiments were performed with both sets of input descriptors. In all the experiments, we followed the cross-subject protocol in which half of the subjects were used for the training phase and the normal were used for testing. The performance of the proposed descriptors was evaluated using ELM and SVM classifiers.

The feature descriptors of the proposed method clearly related with the facial region, where the raw feature takes place. Thus, we experimented with two local patches of size $(20 \times 20 pixels)$; one patch is located on the front region and the other on the face region. The size of the patch is small to keep track of the face region and prevent the patch from biasing toward the background. We selected maximum Max. and minimum Min. values as raw thermal channels, which were extracted from each patch.

In all experiments, we followed the cross subject protocol. In which half of the subjects were used for training phase, and the normal were used for testing.

Results of Experiment 1

In this experiment, local front patch is selected to extract the feature vector. We evaluated the proposed method using NRTLBP and TP1D descriptors along with maximum and minimum values of the raw channels. The recognition rates are reported in Tables 3.4 $-$ 3.9. The results in each table compare the classification rates of NRTLBP and TP1D along with each of the selected raw channel (i.e., with maximum and minimum temperature values). The recognition rates reported in these tables correspond to the following three expression events: normal, pain, and post-pain. We started the experiment by directly providing the raw features of the maximum and minimum temperature values into classifiers separately. The results are reported in Tables 3.4 and 3.5. We also ran the experiment using the NRTLBP descriptor along with both raw channels. The results are shown in Tables 3.6 and 3.7 for ELM and SVM, respectively. The performance of the TP1D descriptor is computed with both raw channels and classifiers. The results of TP1D performance are reported in Tables 3.8 and 3.9. The results shows the ability of the proposed feature descriptors to capture temporal evolution. As indicated in Tables 3.4 and 3.5, the best average recognition rate of ELM is 70% and 65% for Max and Min, respectively . When directly applying Max and Min values into the classifiers, the results show poor feature discrimination of these raw channels, which highlights the challenge in distinguishing event intensity. Tables 3.6 and 3.7 demonstrate the superior performance of the ELM classifier to that of SVM in both Max and Min temperature values. Moreover, the combination of NRTLBP with the maximum temperature value and ELM produced results that are better than those of the combination of NRTLBP with SVM in both raw channels. Furthermore, the accuracy of NRTLBP+ELM is 3.33% better than the accuracy of NRTLBP+SVM using the Max value and 8.33% better than the accuracy of NRTLBP+SVM using the Min value. These results clearly demonstrate the superiority of the proposed NRTLBP representation over the results in Tables 3.4 and 3.5, and suggests that the combination of NRTLBP with ELM is well suited to encode temporal evolution in an assessment-based event representation. Tables 3.8 and 3.9 report the recognition rates for TP1D-based event representations on Max and Min raw channels. The average accuracy of the proposed TP1D+ELM is 8.33% better than the average accuracy of TP1D+SVM using the Max value and 20% better than the average accuracy of TP1D+SVM using the Min value. The proposed TP1D has distinct effects on recognizing events, such as normal and pain from facial intensity, and it obtains the best results along with the combination of ELM and Max value channels.

Table 3.4: Recognition rates when the raw Max. value is directly applied into classifiers

Events	ELM	SVM
normal	65	55
Pain	70	65
Post-pain	75	60

Table 3.5: Recognition rates when the raw Min. value is directly applied into classifiers

Events	ELM	SVM
normal event	60	50
pain	65	65
post-pain	70	75

Results of Experiment 2

For further evaluation, we conducted the experiments when the facial features (i.e., the raw channels) are selected from the local face region with a size of 20×20 pixels . We reported the recognition results when applying the proposed NRTLBP and TP1D descriptors along both maximum and minimum raw thermal channels. ELM and SVM are used to classify thermal features into three categories: normal, pain, and post-pain. The first experiment was conducted by directly providing the raw features of the maximum and minimum values into the classifiers separately. In this case, the results reported in Tables 3.10 and 3.11 compare the classification performances of ELM and SVM for each of the following expressions: normal, pain, and post-pain events. We also ran the experiment by using NRTLBP descriptor along with both raw channels. Tables 3.12 and 3.13 shows the results of ELM and SVM for the NRTLBP feature descriptor along the Max and Min channels. To compare feature descriptors, TP1D is used to measure the performance results on both raw channels. The performance results of TP1D are reported in Tables 3.14 and 3.15.

As indicated in Tables 3.10 and 3.11, ELM and SVM tend to be in line with each other on some levels of events. In the case of using Max value, ELM has the best recognition results of 75% over SVM in pain events, while SVM obtains better re-

Table 3.6: Recognition rates using the proposed NRTLBP descriptor on the Max. raw channel

Events	ELM	SVM
normal event	85	80
pain	100	85
post-pain	80	90

Table 3.7: Recognition rates using the proposed NRTLBP descriptor on the Min. value

Events	ELM	SVM
normal event	75	60
pain	75	65
post-pain	80	80

Table 3.8: Recognition rates using the proposed TP1D descriptor on the Max. value

Events	ELM	SVM
normal event	90	75
Pain	85	90
Post-pain	100	85

Table 3.9: Recognition rates using the proposed TP1D descriptor on the Min value

Events	ELM	SVM
normal event	75	60
Pain	80	55
Post-pain	90	70

Table 3.10: Recognition rates when the raw Max. value is directly applied into classifiers

Events	ELM	SVM
normal event	60	55
pain	75	65
post-pain	70	75

Table 3.11: Recognition rates when the raw Min. value is directly applied into classifiers

Events	ELM	SVM
normal event	70	60
pain	55	65
post-pain	75	80

Table 3.12: Recognition rates using the proposed NRTLBP descriptor on the Max. raw channel

Events	ELM	SVM
normal event	90	80
pain	95	90
post-pain	90	75

sults than ELM in the post-pain event. In the case of the Min value, SVM obtains better results than ELM in the post-pain event. Tables 3.12 and 3.13 illustrate that NRTLBP+ELM+Max have distinct effects on recognizing all events of facial intensities. For example, an ELM has a stable recognition rate of 95.00% in pain versus 90.00% in normal event. In general, ELM with NRTLBP+Max has the best overall classification performance. Tables 3.14 and 3.15 reports the recognition rates for TP1D representations on Max and Min raw channels. The average accuracy of TP1D+ELM is better than that of TP1D+SVM for normal and pain events using Max value of the raw channel. Table 3.15 shows the performance of TP1D+ELM over TP1D+SVM for the normal and post-pain events using the Min value of the raw channel. The proposed TP1D has distinct effects on recognizing events from facial intensity. Moreover, the classification accuracy of TP1D using ELM+Max is significantly better than that which uses SVM+Max.

Table 3.13: Recognition rates using the proposed NRTLBP descriptor on the Min. raw channel

Events	ELM	SVM
normal event	75	80
pain	65	65
post-pain	70	65

Table 3.14: Recognition rates using the proposed TP1D descriptor on the Max. raw channel

Events	ELM	SVM
normal event	100	85
pain	90	75
post-pain	85	80

Table 3.15: Recognition rates using the proposed TP1D descriptor on the Min. raw channel

Events	ELM	SVM
normal event	80	75
pain	75	80
post-pain	85	70

In summery, the results presented here indicate that the thermal signature representation in NIC may be included into the future neonatal monitoring modalities for medical assessment and vital sign diagnostic. At currently, the results acquired during measurement are not fully categorized and need more reliable measurement protocols. Additionally, the neonatal thermal measurement are correlated with other physical operators, such as opening and closing incubator, respiration mask of infant, etc. However, behavior discrimination in neonatal monitoring remains a challenge due to a complex interactions e.g. face masks or prongs, mechanical ventilation, head rotation, motion artifacts, etc.

Furthermore, whereas thermal imaging has been mainly applied to object detection in night monitoring, in this work thermography was shown to allow non-contact event monitoring in CMinICU. Physically, the work is based on changes the temperature intensity of facial region, induced by pain response in dedicated ROIs. Until now the method achieves better, and the presented results are preliminary and need further studies in a larger number of neonates and under a variety of care setups. The analysis method needs further improvement, such as automatic local ROI definition. Furthermore, the thermal monitoring may be also considered as a first step to evaluate non-invasive behavior of premature infants.

3.3 Conclusion

In this chapter, two thermal feature descriptors were proposed to represent temporal evolution of raw thermal signals. These descriptors efficiently represent the local skin underlying temperature profiles centered on the tip of the front and face areas. We used a local patch defined over the ROIface to extract raw thermal channels. We adopted two descriptors to encode the temporal information over the time sequence of raw thermal signals to provide distinctive information about the signal attributes. Moreover, the descriptors mapped the raw thermal features into local temporal binary codes and topological persistence space, that is, they allowed us to use efficient temporal feature descriptors to extract the most discriminative pattern sequence. We used ELM and SVM as the classifiers for event recognition. We evaluated the discrimination power of the descriptors on the task of premature infant-based real event recognition from premature infant data captured by thermal video. We demonstrated via experiments that the proposed representation achieved the best results for event discrimination.

We used non-redundant local binary features and topological computation to model the temporal variation of a manually sampled signal over a local area. However, each event is usually characterized by a set of factors. Hence, we are planning to explore new schemes to efficiently identify the set of attributes that differentiate the local attributes of the signal. This study is one of the first attempts to apply the concepts of local neighbor interaction and topological computation to actual temporal evolution problems.

The results of this study are promising and suggest that the proposed feature representation algorithms could prove useful in condition monitoring assessment. Furthermore, this method may be an effective quantitative technique to demonstrate the pain response pattern in premature infant infants. This possibly can give information about the depth and the frequency of each event to get an early sign of changes in the infant behavior.

Part II

Human Action Recognition from 3D Body Skeleton

CHAPTER 4

3D-POSTURE RECOGNITION USING JOINT ANGLE REPRESENTATION

Contents

In this chapter, we propose a method for posture-based human action recognition. In the proposed work, 3D locations of joints from a Skeleton information are considered as initial inputs to our method. Skeletal joint positions are first projected into hip area of the body skeleton, and simple relation between coordinates vector is used to describe the 3D body coordinates. We perform the representation of human postures by selecting 7 primitive joint positions, which generates a compact feature called Joint angles. To make the skeletal joint representation robust against minor posture variation, angles between joint are cast into orthogonal planes of xy and zy, respectively. The vector of joint angle features is quantized through unsupervised clustering into k pose vocabularies. Then encoding temporal joints-angle features into discrete symbols is performed to generate Hidden Markov Model HMM for each action. We recognize individual human action using generated HMM. Experimental evaluation shows that this approach outperforms state-of-the-art action recognition algorithms on depth videos.

4.1 Proposed approach

4.1.1 Body skeleton representation

In this section we describe the human poses representation and joints position estimation from skeleton model. This kind of representation consists of 3D joints coordinates of a basic body structure, which consisting of 20 skeletal joints as shown. Recent release of RGBD system offers better solution for the estimation of the 3D joint positions. An example in the Figure 4.1 demonstrates the result of depth map and the 3D skeletal joints according to algorithm of [Bloom et al., 2012] which proposed to extract 3D body joint locations from a depth map. The algorithm of [Shotton et al., 2011] is used to estimate pose locations of skeletal joints. Starting with a set of 20 joints coordinates in a 3D space, we compute a set of features to form the representation of postures. Among the 20 joints, 7 primitive joints coordinates are selected to describe geometrical relations between joints. The category of primitive joints offers redundancy reduction to the resulting representation. Most importantly, primitive joints achieve view invariance to the resulting pose representation, by aligning the Cartesian coordinates with the reference direction of the person. Moreover, we propose an efficient and view-invariant representation of postures using 7 skeletal joints, including *L/R hand, L/R feet, L/R hip, and hip center*.

The hip center is considered as the center of coordinate system, and the horizontal direction is defined according to left hip and right hip junction. The remaining 4 skeletal locations are used for poses joint angles descriptor.

Action coordinates for skeletal joints

The output from 3D sensor system contains the most useful raw information about the motion sequences, such as the depth image(D), body part relations of the joints, and relative angles.

In order to make the 3D joints locations invariant to sensor parameters. We thus necessarily need to register the hole skeleton body into a common coordinate system, along the action sequence.

Therefore, to aligned the body skeleton into the reference coordinate system. we take the hip center hc as the origin of the reference coordinate system, and use its coordinates as the common basis, and define the horizontal reference vector ρ

to be the vector from the left hip center to the right hip center projected on the horizontal plane as depicted in Figure 4.2. In this work, the subject coordinates comprises the following three orthogonal vectors $\{\rho, \gamma, \beta\}$that are identified as

$$\begin{array}{ll} \rho = \frac{\mathbf{j_1} - \mathbf{j_2}}{\|\mathbf{j_1} - \mathbf{j_2}\|}, & \gamma = \frac{\rho \times \mathbf{u}}{\|\rho \times \mathbf{u}} \\ \mathbf{u} = \frac{\mathbf{j_3} - \mathbf{j_2}}{\|\mathbf{j_3} - \mathbf{j_2}\|}, & \beta = \rho \times \gamma \end{array} . \tag{4.1}$$

Where j_i hip center, R/L hip joints, respectively, $||.||$ denotes the norm of a vector and $\times$ denotes the cross product of two vectors. As an illustrative example the aligning of subject's' coordinates procedure is depicted in Figure 4.1

4.1.2 Features description

In this context, we choose to represent skeleton body in terms of the angles between joints, which showed to be more accurate than using e.g. directly the joints 'coordinates. In order to compute a compact features, the aforementioned angles are extracted in the orthogonal planes. Moreover, all angles are computed using the hip-center joint as reference, i.e. the origin of the coordinate system is placed at the hip-center joint coordinate.

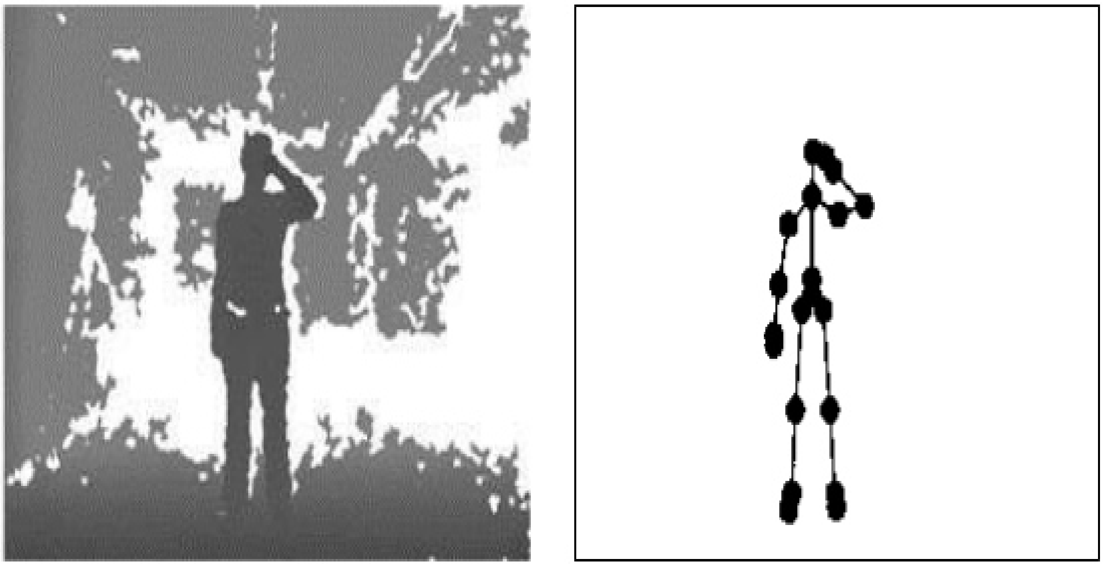

Figure 4.1: (Left)Depth map image. (right) Skeletal joints positions proposed by Bloom et. al. [Bloom et al., 2012]

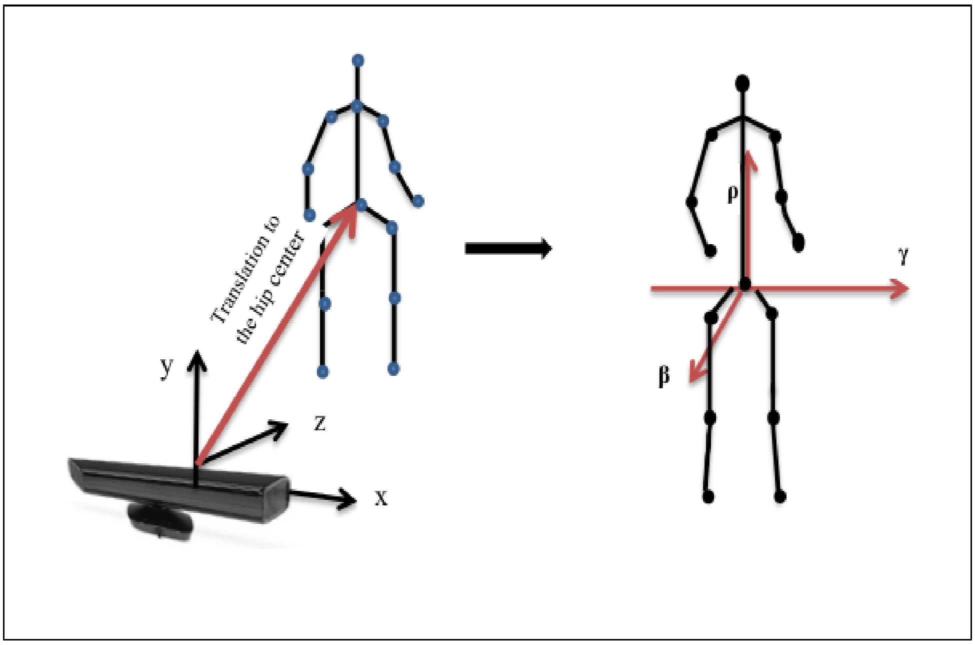

Figure 4.2: 3-dimensional coordinates corresponding to a human body skeleton

For computing the proposed action representation, only a primitive set of the supported joints which defined in above section is used. To this end, only the angles between,*hand left/foot left, hand left/hand right, hand right/foot right, and foot right/foot left respectively* are extracted as shown in Figure 4.3. The angles between joints are sampled from orthogonal planes, XY and ZY planes with respect to the origin. In each plane, four angles are quantified using the trigonometric function. The skeletal joint-based pose representation is computed by casting the 8 angles into the corresponding feature vector. Moreover,The final features vector includes eight joints angles of $F_t = \{\theta_1, \theta_2, \cdots, \theta_8\}$ at each pose instant t.

4.1.3 HMM for action recognition

To apply HMMs to problem of human action recognition, the video frames $V = I_1, I_2, \ldots, I_T$ are transformed into symbols sequences O. The transformation is done through the learning and recognition phases. From each video frames, a feature vector $f_i \in R, \{(i = 1, 2, \ldots .T)\ T$, number of the frames$\}$ is extracted, and f_i is assigned to a symbol v_j chosen from the set of symbols V. In order to specify the observation symbols, we perform clustering of feature vector into k clusters using K-means algorithm. Then each pose instance is represented as a single number of a code word. In this way, we collect for each action a sequence of the visual words. The obtained symbol sequences are used to train HMMs to

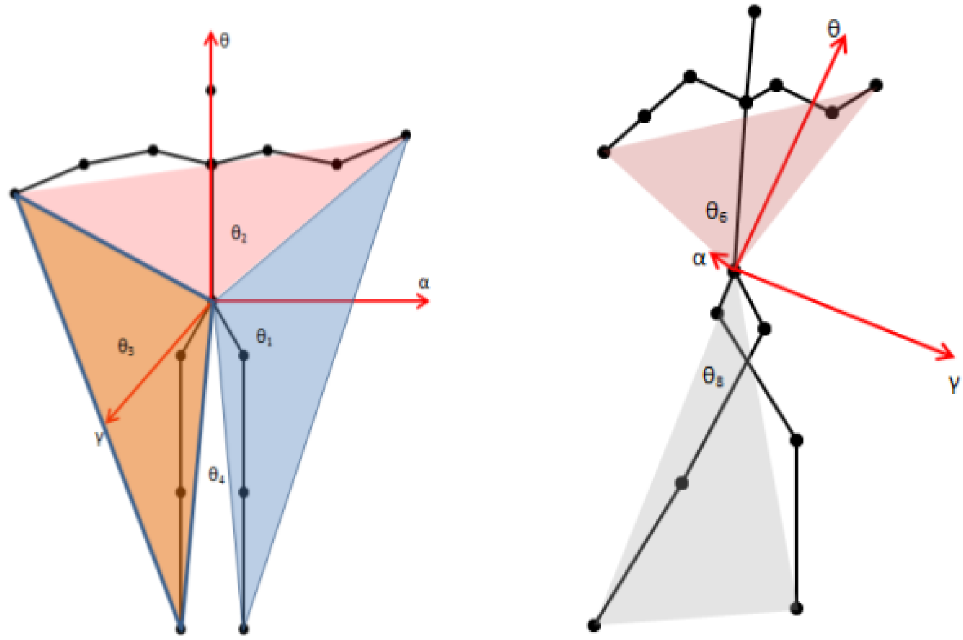

Figure 4.3: 3-dimensional coordinates corresponding to a human body skeleton

learn the proper model for each activity. For recognition of a test activity, the obtained observation symbol sequence $O = O_1, O_2, \ldots .O_N$ is used to determine the appropriate human action HMM from all trained HMMs.

HMMs, which have been recently applied with particular success to speech recognition, are a kind of stochastic state transit model [Rabiner, 1989]. HMM is using observation sequence to determine the hidden states. We suppose $O = O_1, O_2, \ldots .O_N$ as the observation of the stochastic sequence. HMM with $NS = s_1, s_2, .., s_N$ state is specified by the triplet $\beta = A, B, \pi$ of parameters. More specifically, assume we denote by S_t the state at time instance t. The state transition probability matrix, used to describe the state transition between probability is given:

$$\mathbf{A} = \{a_i^j = P_r(s_{t+1} = q_j | s_t = q_i)\}. \tag{4.2}$$

Where, a_i^j is the probability of transiting from state q_i to stateq_j.

The matrix of observation probabilities, Used to describe observed values $b_j(k)$ of output symbol v_n at state q_j is

$$\mathbf{B} = P_r(v_n | s_t = q_i). \tag{4.3}$$

And the initial state distribution vector π is

$$\pi = \{\pi_i = P_r(s_1 = q_i)\}. \tag{4.4}$$

In training phase, we create a single HMM model for each of the actions. Then for an action sequence $V = v_1, v_2, .., v_T$ we calculate the model $P_r(V|\beta)$ of the observation sequence using the forward algorithm. To this end, the action can be classified as the sequence which has the largest posterior probability as:

$$\mathbf{L} = \arg\max_i P_r(O|\beta_i). \tag{4.5}$$

Where i indicates the likelihood of test sequence for the ith HMM.

4.2 Experimental results

We evaluate our proposed method on different public datasets: MSR Action3D, and gaming3D-Action. For each dataset, we extensively compare the State-of-the-art skeleton-based methods to our approach. Note that we only used the information from the skeleton for action recognition in our algorithm. The set of clusters and number of state were fixed to K=80, and N=6. Cross subject testing was used in the recognition system, i.e. half of the subjects were used for training HMMS and the rest of the subjects were used for testing.

The proposed algorithm is tested on the MSR Action dataset using cross subject. As originally proposed [Wanqing et al., 2010]the dataset was further divided into subsets AS1;AS2 and AS3, each consisting of 8 actions see Table 1.1. We performed recognition on each subset separately and all the results were averaged over these subsets. Each test is repeated 10 times, and the average performance is reported. We compare the performance with state-of-the-arts methods.
Table 4.1 reports the recognition rates of our method on MSR-Action3D dataset. The recognition rates in the last row are the average of the recognition rates for the three subsets AS1, AS2 and AS3. Table 4.2 reports the competitive result of the proposed approach along with the corresponding accuracies of methods that focus on skeleton joints action representation or depth information. It is worth to note that our method outperforms the majority of these methods. Specifically, it outperforms the state-of-the-arts [Ofli et al., 2012, Lv and Nevatia, 2006b,

Wanqing et al., 2010] by 29.4% , 13.36 %, and 1.76 %, respectively.

Table 4.1: Recognition rate of proposed method on the MSR Action dataset

Action Subset	Accuracy
AS1CrSub	86.30
AS2CrSub	65.40
AS3CrSub	77.70
Overall	76.46

Table 4.2: Compare recognition rate (%) of proposed method with the state-of-the-arts results, on the MSR Action 3D dataset.

Methods	Overall
Joint angles + SIMJ (Ofli et al. [Ofli et al., 2012])	47.06
Hidden Markov Model [Lv and Nevatia, 2006b]	63
Bag of 3D points (Li et al. [Wanqing et al., 2010])	74.70
Histogram of 3D Joints (Xia et al. [Xia et al., 2012]	78.97
Our method	76.46

Table 4.3: Recognition accuracy on G3D dataset using skeleton joint

Action category	Bloom et al. [Bloom et al., 2012]	Our method
Fighting	70.46 %	79.84 %
Golf	83.37 %	100 %
Tennis	56.44 %	78.66 %
First Person Shooter FPS	53.57 %	54.10 %
Drive car	84.24 %	81.34 %
Miscellaneous Misc	78.21 %	89.40 %
Overall	71.04 %	80.55 %

We also tested the proposed method on the public G3D Action database that was released for real time gaming-based dataset. As originally proposed in [Bloom et al., 2012]. the actions are divided into 7 actions categories. We used the same parameter settings as previously. Each test is repeated 10 times, and the average performance is shown in Table 4.2. It can be noted that the proposed method achieves considerably higher recognition rates than Bloom et.al. on some actions category. In particular, the proposed method actually achieves much higher recognition accuracy on Fighting, Golf, Tennis, and Misc categories. While have recognition rate is slightly lower on Drive Car action group. However, from Table 4.3, we can see that the proposed work perform better on 4 of the 6 action group.

In summery, The joint angles based human skeleton representation works better than the approaches based on direct joint position. It seems natural that the orthogonal partition information is very important for action recognition, and often more important for view invariant than the relative distance or joint velocity information. Moreover, angles between joints in orthogonal planes encoding works better than the joint position or distance between joint features approach. The main cause why the angle between joint encoding works better than, for example joint distance features approach, may be the loss of temporal information by the joint distance features approach when doing the complex action and poor assignment of local features to visual words.

4.3 Conclusion

This chapter presents a novel 3D skeleton representation framework for 3D pose based action recognition. According to the proposed scheme, skeletal raw data are projected into body coordinate system and five joints of human skeletal are chosen.. The 3D pose representation of human action is represented by computing the angels between joints in orthogonal planes, which constitutes view invariant description of 3D human poses. Angels Feature vectors are then concatenated in order to characterizes the prototype of the action. A set of code words are built by clustering the large collection of feature vector. Discrete HMMs are learned and used to classify sequential poses into action types.

We demonstrated the power of our methodology by obtaining state of the art results on recent, challenging benchmarks for action and gaming recognition. The evaluation procedure summarizes that the proposed method achieves reasonable performance on dataset includes both view invariant and temporal nature challenging. Since despite its simplicity, the proposed method provides compact 3D pose description. Additionally provides a flexibility of incorporates more skeleton joints in order to increase the reliability of accuracy in several applications.

Chapter 5

3-D Skeleton joints-based action recognition using covariance descriptors on discrete spherical harmonics transform

Contents

This chapter attempts to address the skeletal-joints representation problem in an explicit model. In this model, a novel feature descriptor is used based on the Spherical Harmonic Transform (SHT) of temporally local joints and the covariance coefficients. The main objective of our approach is based on the calculation of the SHT of spherical angles of local joints to explicitly model the displacement of each individual joint. Unlike the traditional works that consider the spatial-relation between individuals joints. While the present study is related to recent approaches in skeleton descriptor, it capitalizes on a new feature space, which was not considered in these earlier studies.

Let a spherical coordinates of skeleton joint Ji denoted by (θ, ϕ), the model of temporal evolution ofJi can be represented using spherical harmonic of θ and ϕ orientation respectively. Then, to handle frame length variations, for each action category, we introduce the covariance technique to compute the covariance coefficients of each SHs matrix. Collecting the computed covariance coefficients of all selected local joints forms the skeleton features representation for an action sequence.
Finally, we present an extensive evaluation of the proposed skeleton-based descriptor with the covariance encoding and the Extreme Learning Machine ELM on four various 3D action recognition datasets. We show that the proposed descriptor always achieves better results than the existing skeleton-based state-of-the art algorithms.

5.1 Introduction

In chapter 4, we present simple skeleton body joint representation for action recognition in RGBD video, which model the skeletal sequence as an angles between four primitive joints. the proposed representation can encode view invariant by extracting an angels feature vector via orthogonal plane partitioning. However, despite the proposed method work better, the proposed method is not good enough to encodes the temporal nature, and joint displacement in good way. We observe that typically the joint movement based representation can provides motion information about complex action which exhibit temporal variation. As skeleton-based human action recognition techniques have shown to achieve good results, we believe that more discriminative representation for modeling human body skeleton could be proposed. Therefor, in this chapter we primary focus on explicitly modeling the body skeleton joint by projecting the spherical orientation of the joint into 3D Fourier harmonics basis.

Most the existing skeleton-based approaches are focused on features based on the distance information between the joints and features based on the 3D coordinates of the joints. These methods directly fed theses features into recognition system. However, the temporal dependency and relation between individual features are not considered by these techniques. Therefore the aforementioned methods might not enough provides discrimination ability to recognize complex actions. Different from the existing techniques, we introduce a novel human body skeleton description for action recognition. Moreover, we project the Spherical angular vector of local joint into Spherical Harmonics SH (i.e. 2-Sphere) to explicit modeling the human body skeleton. In order to encode the structural information between joints, we compute a covariance representation on SHs called CSHs. We present an extensive evaluation of the proposed approach on four various state-of-the-art datasets. We show that the CSHs outperform state-of-the-arts skeleton-based action recognition.

5.2 Proposed approach

5.2.1 Body Coordinates Projection

Human poses are represented by a skeletal structure composed of 20 joints. Such a representation is generated with the RGBD sensor as an example. Each joint is represented by its 3D position in the real world. Fig. 5.1 shows a sample skeletal structure corresponds to the 20 joints. Essential joints, including { hands, feet, elbows, knees, and root}, are found in all skeletons and are regarded as the key joints in translating the skeletal coordinates.

Several techniques for joint coordinate transformation incorporate various relational features with skeletal data or rely on the modified Cartesian and spherical coordinates. In [Müller et al., 2005], the authors adopted projections of velocity vectors on the plane defined by the shoulder and hip points. The torso principal component analysis (PCA) frame denotes another transformation and was recently proposed in [Raptis et al., 2011]. This method is based on the assumption that torso joints (shoulders and hips) rarely move independently; thus, the torso can presumably be a rigid body. The authors proved that the orthonormal basis of the corresponding torso joints can be determined by conducting PCA projection on the coordinates of the torso area. Directly modified joint coordinate methods are explored in many works as well. [**?**] modified these methods by aligning spherical coordinates with the specific direction of a person. Furthermore, these researchers defined the center of the spherical coordinates as the hip center joint. The horizontal reference vector is considered the direction from the left hip center to the right hip center as projected on the horizontal plane, and the azimuth angular vector is the vector that is perpendicular to the ground plane and passes

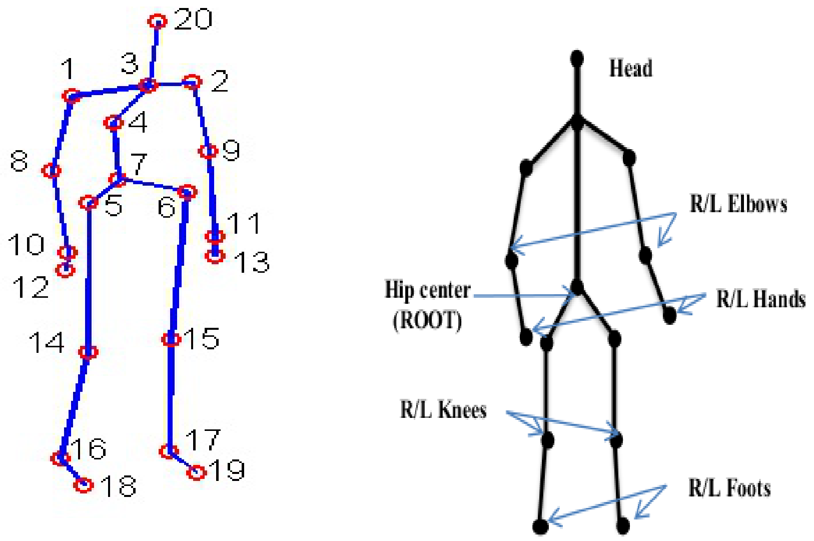

Figure 5.1: Skeleton model: Left Skeleton model of 20 joints, Right selected joints for pose representation

through the coordinate center.

However,to transform the joints into the body coordinate system, we select the hip center as the reference joint, apply its coordinates as the common basis, and transform all the other skeletons in the sequence to this joint.

As per this work, the origin of spherical coordinates is positioned at the hip center by subtracting the coordinates of the root from all joints. To realize the view invariance of skeletons, we utilize the rotation matrix to rotate these skeletons such that the vector from the left hip to the right hip is parallel to the horizontal coordinate system, as illustrated in Figure 4.2. We also normalize the angular directions of the skeleton to be scale invariant. The new coordinate system can approximate the trajectories of body joints and depends only on the directionality of the joints around the hip area, that is, the hip center and the left/right hips.

5.2.2 Spherical angular estimation

For efficient pose representation that satisfactorily handles the view invariance and independence from the relative position of the subject to the sensor, we represent a pose in terms of the angles of individual skeletal joints as expressed in the proposed coordinate system. This approach is more discriminative than directly applying the normalized joint coordinates. To compute for a compact description, the aforementioned angles are estimated in the spherical coordinate system as follows:

$$
\begin{aligned}
\theta(\mathbf{t}) &= \arctan\left(\tfrac{y}{x}\right) \\
\phi(\mathbf{t}) &= \arccos\left(\tfrac{z}{\sqrt{(x^2+y^2+z^2)}}\right)
\end{aligned}
\tag{5.1}
$$

where t is the frame index, and $\theta_i\ and \phi_i$ are the estimated spherical angles. Figure. 5.1 explains the selected skeletal joints in this context. Only a subset of the primitive joints is used because the trajectories of certain joints are close to one another and are thus redundant in describing the configuration of body parts. Otherwise, these trajectories contain noisy information. To this end, only the joints that are presumably the most appropriate, that is, those that correspond to the upper and lower body limbs, are considered. These joints are the *right/left elbows, right/left hands, right/left knees, right/left feet, and head.*

Therefore, each pose is represented by a raw vector that consists of spherical angles (θ, ϕ). The right panel in Fig. 5.2 indicates the spherical orientation of each selected joint. The obtained spherical angles may improve the performance of the proposed method because they detect characteristic motion patterns among individual joints. Rotation invariance can be achieved explicitly by considering spherical directions instead of an absolute joint position. Carefully estimating the obtained spherical directions resolves significant ambiguities in the execution of action pairs, such as punching and kicking, hand waving, and golf chipping.

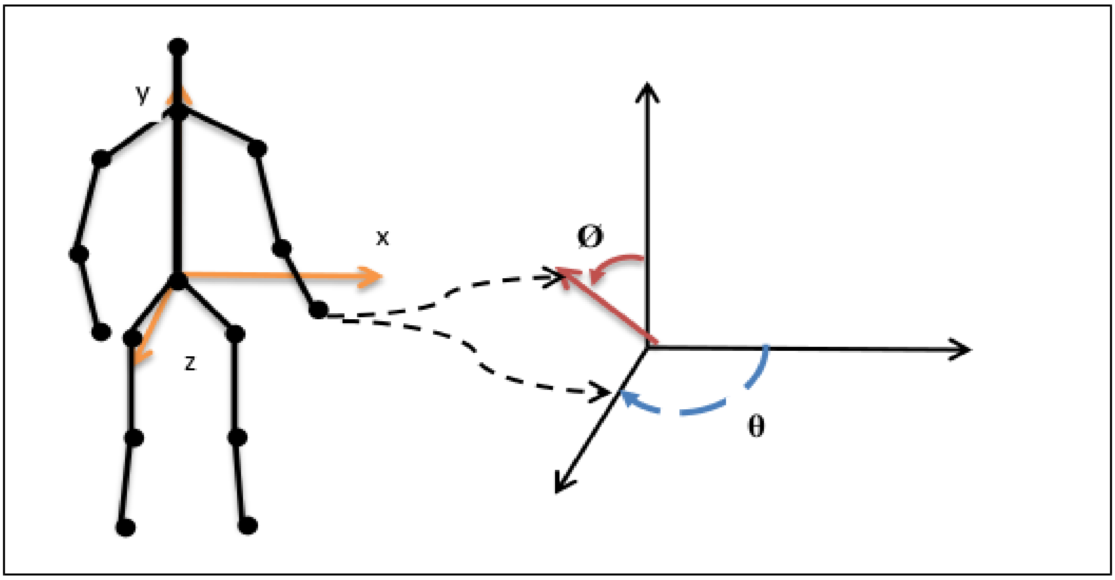

Figure 5.2: Euler angles of selected joints expressed in the 3-D Spherical coordinates

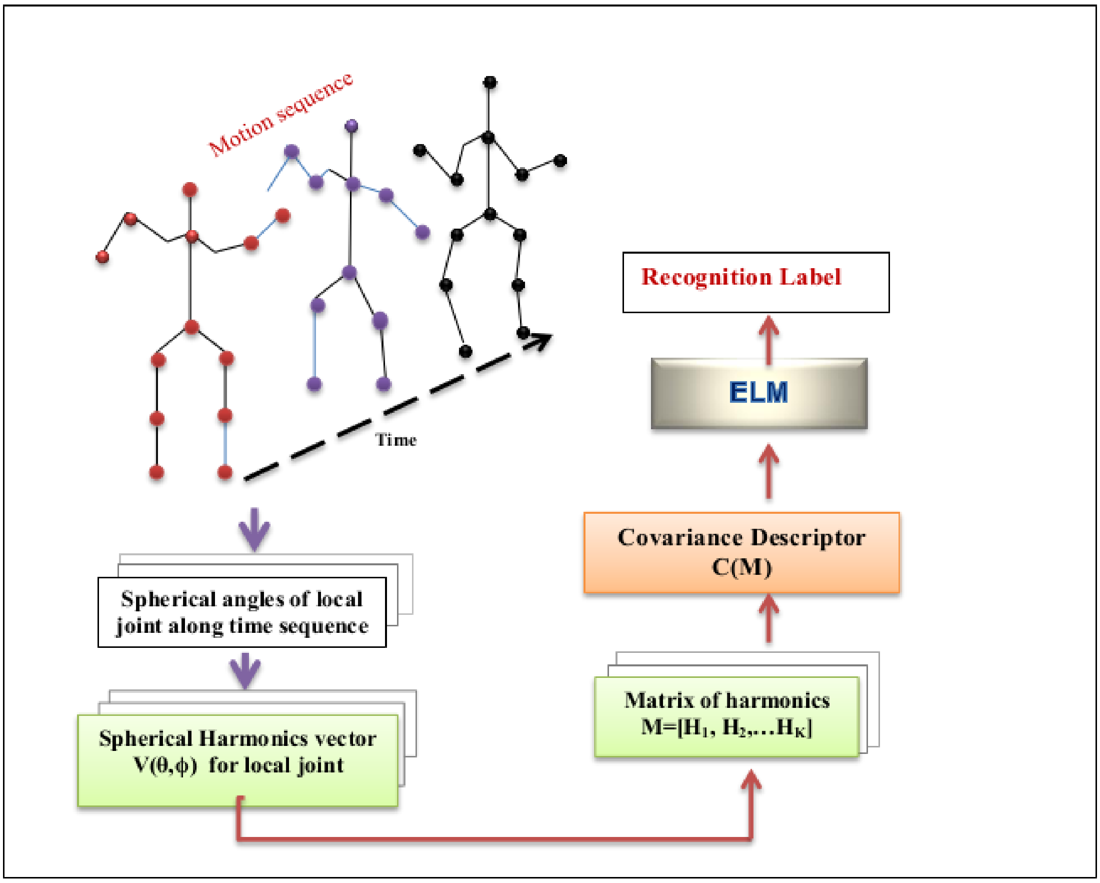

Figure 5.3: Overview of the calculation process of the proposed method. Firstly, we extract temporal spherical orientations of the joint . Then we represent theses angels using SHS. Then we use covariance property to build the action descriptor on SHs. We apply ELM for recognition task

5.2.3 3D pose descriptor

A good descriptor should capture both static poses and joint kinematics at a given moment in time to realize a robust representation that counters minor joint location errors. However, most methods recognize motion by directly classifying the features extracted based on joint position [Hussein et al., 2013], pairwise distance [Antônio et al., 2012], differences in joint position [Yang and Tian, 2012], and body part segments [Evangelidis et al., 2014]. These approaches aim to model the motion of either individual joints or the combinations of joints according to the aforementioned features. A compact and efficient skeleton description has been provided as an explicit model. Such methods straightforwardly model joint information in appropriate spaces. In [Theodorakopoulos et al., 2014], a skeleton was represented via sparse coding in a dissimilarity space. An alterna-

tive path was proposed by [Vemulapalli et al., 2014] in which skeletal joints were modeled as a point in the Lie group (special Euclidean space). 3D human actions were represented in [Devanne et al., 2013] by the spatio-temporal motion trajectories of pose vectors. These trajectories were represented as curves in the Riemannian manifold of an open-curve shape space to model the dynamics of temporal variations in poses.
The overview of the calculation process of the proposed CSHs descriptor for human skeleton representation is illustrated in Figure 5.3. However, we are given a vector of spherical orientation (or Euler as interchangeable) which belong to individual joint, and our goal is to create its compact descriptor in SHs.

Skeleton joint representation using SHs

SHs are versions of trigonometric functions for the Fourier expansion on the unit sphere s^2. The properties of spherical modeling in terms of these harmonics are naturally observed during analysis in the fields of theoretical physics, geoscience, and astrophysics, among others. In this section, we review SHs.
SHs are an extension of Fourier techniques to three dimensions and are particularly well suited for modeling shapes from such data. These harmonics are applied to related problems in computer vision and in 3D model retrieval [Bustos et al., 2005, Saupe and Vranić, 2001], rotation invariance, descriptor-based 3D shapes [Vranic, 2003], and face recognition under an unknown lighting constraint [**?**] and [Romdhani et al., 2006]. The rich material in [Lebedev N., 1972] provides a general introduction to SHT and presents classical tools of SHs.

Let $(r, \theta, \phi) : r \in R^+, \theta \in [0, 2\pi], \phi \in [0, \pi]$ be the spherical coordinates and $f(\theta, \phi)$ be the homogeneous harmonic functions on R^3. In the current study, we aim to determine the homogeneous solutions of the Laplace equation $\nabla^2 f = 0$ in spherical coordinates. Likewise, we intend to explain how these solutions correspond to the decomposition of eigenfunctions in space $L^2(S^2)$, $S^2 = (x, y, z) \in R^3$. In this case, SH generalizes the Fourier series to two spheres by projecting the square-integrable function S^2 onto the Hilbert space $L^2(S^2)$.

Firstly, for the following spherical coordinates:

$$\begin{aligned} x &= r\sin\theta\cos\phi \\ y &= r\sin\theta\sin\phi \\ z &= r\cos\theta. \end{aligned} \tag{5.2}$$

The Laplacian of a harmonic function using angular version is given by:

$$\mathbf{\Delta}_{\mathbf{S}^2} f = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial f}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 f}{\partial\phi^2}. \tag{5.3}$$

The final solution of Laplacian in R^3 (due to space limitation, the detailed solution is no longer provided) is a set of Legendre function and eigenfunctions expressed as follows:

$$f(\theta,\phi) = K(P_l^m(cos\theta))(\exp(jm\phi)). \tag{5.4}$$

Where K is a constant. The first term in Equation 5.4 is a set of Legendre polynomials, and the second term is the eigenfunctions of the Laplacian on a sphere with an eigenvalue of $l(l+1)$. The notation of the preceding equation represents the SHs in complex form. In this context, we adopt the notion of real SHs with the degree of l and order of $m > 0$. Thus, we set

$$y_n^m(\theta,\phi) = \sqrt{(2)}K_n^m\cos(m\phi)P_n^m(\cos\theta). \tag{5.5}$$

where $P_l^m cos(\theta)$ are the associated Legendre polynomials of degree l and order m, defined by the differential equation as

$$P_l^m = \frac{-1^m}{(2^l l,\ !)}(1+x^2)^{\frac{m}{2}}\frac{(d^{l+m})}{(dx^{l+m})}(x^2-1)^l. \tag{5.6}$$

And the trem K_l^m is a normalization constant, equal to

$$K_l^m = \sqrt{\left(\frac{(l+1}{4m}\right)\frac{(l-|m|)!}{(l+|m|)!}}. \tag{5.7}$$

The author in [Lebedev N., 1972] specify that any function of the form $f(\theta,\phi)$ can be represented by a set of expansion coefficients on the unit sphere. Complete harmonic basis functions are indexed by two integer constants (i.e., the degree l

and the order m).

The sampling frequencies of the basis functions over the unit sphere are defined by the values of the order $-l \leq m \leq l$. $2l+1$ bases are detected in general. Visual representations of the real SHs in the azimuth and elevation directions are displayed in Figure. 5.4 as an illustration. In this figure, the blue portions represent positive harmonic functions, and the red portions depict the negative ones. The distance of the surface from the origin indicates the value of P_l^m in the angular direction (θ, ϕ).

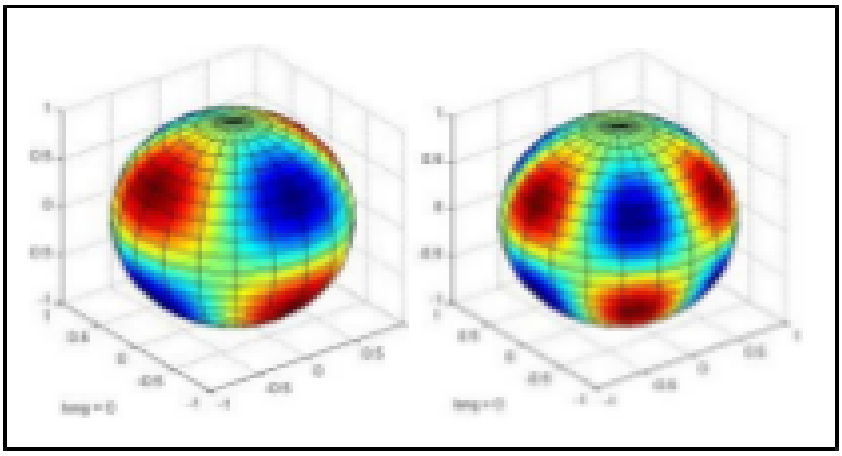

Figure 5.4: Visual representations of the real spherical harmonics . (Right) l=3, m=2.(Left) l=4, m=3

The above definitions typically explain the general solution of laplacian on the angular version. Therefore, to project the spherical angular into the harmonics basis, we decompose the $f(\theta, \phi)$ using discrete SHs. For every local joint in body skeleton, we extract a vector of angular directions(θ_k, ϕ_k), (k: joint index) along time sequence. Thus, we map this vector into a basis functions as:

$$x(\theta, \phi) = \sum_{l=0}^{L\,\max} \sum_{m=-l}^{l} f_l^m Y_l^m(\theta, \phi). \tag{5.8}$$

where L_{max} is a user-defined maximum frequency and f_l^m denotes the expansion coefficients, which are calculated with

$$f_l^m = \frac{4\pi}{n} \sum_{k=0}^{n-1} x(\theta_k, \phi_k)\; Y_l^m(\theta_k, \phi_k), \tag{5.9}$$

the real parts $Y_l^m(.)$ of spherical harmonic are defined as

$$Y_l^m(\theta,\phi) = \begin{cases} \sqrt{2K_l^m}\cos(m\phi)P_l^m(x) & , \ m > 0 \\ \sqrt{2K_l^m}\sin(|m|\phi)P_l^{|m|}(x) & m < 0 \end{cases} \tag{5.10}$$

The equation (5.8) has two fundamental solution, real harmonics spanned by $\cos(m\phi)$ and Legendre polynomials $P_l^m$of degree m. Our demonstrations have shown how a basis of SHs can be computed entirely from $2n+1$ systems of linear equations. In the other hand, the set of solutions in equation (5.8) can be intuitively approximated by the distribution of the positive and negative coefficients on the spherical surface. The discriminative coefficients are distributed according to the frequency band and degree parameters. Figure 5.5 demonstrate practical examples of higher order SHs basis functions decomposition.
Finally, for each individual joint we define its SHs as a 2D matrix. Moreover,the N elements of spherical angels of individual joint form the $N * N$ SHs matrix.

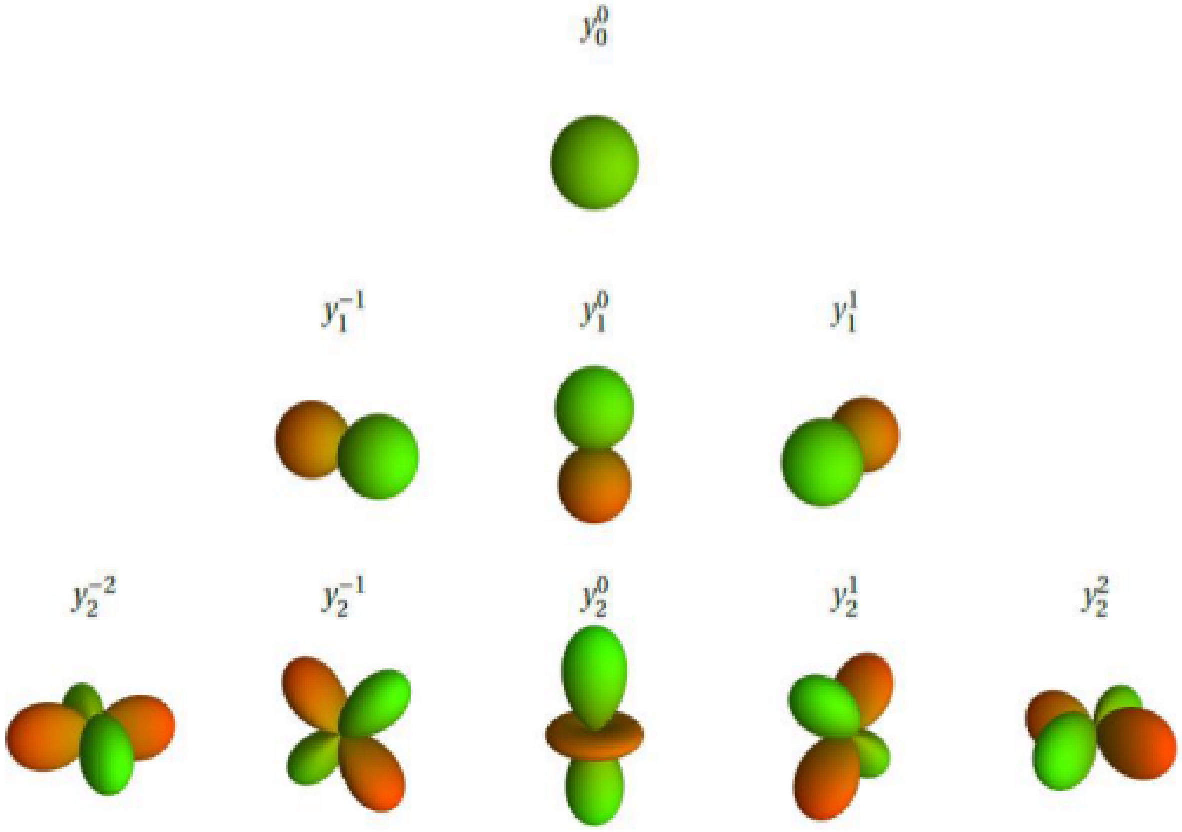

Figure 5.5: Plots of the higher order real-valued spherical harmonic basis functions. Green indicates positive values and red indicates negative values.

Covariance descriptor on SHs

Regardless of the skeleton structure being used, temporal sequence discrimination into different action classes is a difficult task due to challenges like frame numbers variations in each action, and temporal joints dependency. To address these problems for each action class, we propose a highly discriminative 3D pose descriptor. Particularly, we introduce a novel skeleton-joints descriptor that is based on finding the covariance coefficients on the spherical harmonics of local joints. We sample these coefficients over the time of the action sequence.
The idea of covariance descriptor was first adopted by [Tuzel et al., 2006] as a region descriptor of an image and texture-based classification. The idea of spatiotemporal patch-based covariance descriptor is recently introduced as an action recognition framework [Andres et al., 2013, Tuzel et al., 2008] . In our work, we compute the spatiotemporal covariance coefficients between local joints elements which extracted along the time sequence. The overview of the calculation process of the covariance descriptor for a SHs vectors is presented in Figure 5.6.

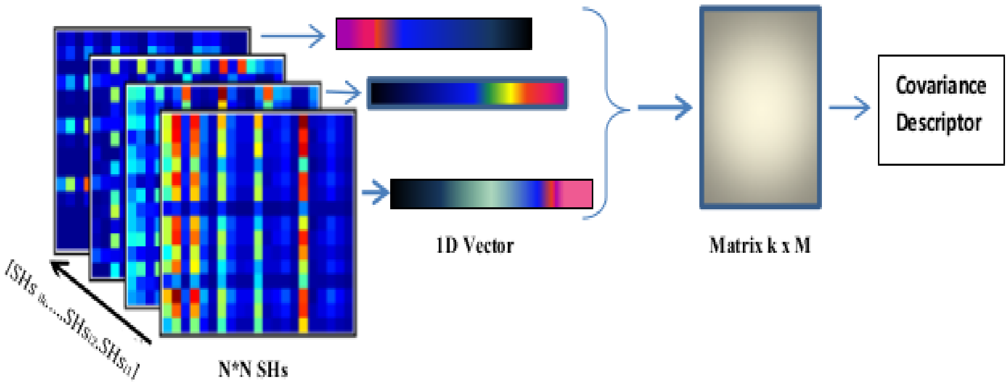

Figure 5.6: process of Covariance Descriptor calculation

Suppose we have the entire skeleton structure is represented by Q joints, and the action is performed over T time sequence (frame). Let H denote harmonics data matrix of a set of spherical harmonics $\{h_1, \ldots ..h_n\}$. Because sets of related spherical harmonics of Q joints are considered for whole action, the 2-D SHs h_i of length $m = v \times u$ is expressed in column vector i.e. $= vect(h)$. Thus, the harmonic data H is an$M \times Q$ matrix, and defined as $\mathbf{H} = \{\mathbf{h_1}, ..., \mathbf{h_Q}\}$ where typically,$M > Q$ with fixed Q. Having obtained the harmonic data matrix$H$, the covariance elements over the sequence T is given :

$$C(H) = \frac{1}{T-1}\sum_{t=1}^{T}(H-\bar{H})(H-\bar{H}). \tag{5.11}$$

Where $\bar{H}$ is the sample mean of H.

In our case, we sample the lower part elements of the covariance matrix $C(.)$. Thus, the length of the descriptor is $Q(Q+1)/2$. Where Q is the number of skeleton joints used to represent the action sequence. The obtained feature vector represent the final features of the action sequence.
Once the descriptors are calculated in a video sequence, we use them to represent this video sequence. Finally, we apply the ELM to classify video representations into action categories.

5.3 Experiments

In this section, we present an evaluation, comparison, and analysis of the proposed method. The experiments are performed on 4 state-of-the-art action recognition datasets. These datasets are: MSR-Action3D dataset [Wanqing et al., 2010], UTKinect-Action dataset [Xia et al., 2012], Florence3D-Action dataset[Seidenari et al., 2013] and gaming G3D dataset [Bloom et al., 2012]. In all experiments, we used a ELM classifier with the covariance descriptor.
For MSR-Action3D dataset, the protocol of cross subject test setting was used similar to [Wanqing et al., 2010]. We further divided the dataset into subsets AS1, AS2 and AS3 each consisting of 8 sub-actions. The recognition task was performed on each subset separately and we averaged the results. For the remaining data sets, we divide each dataset into half of the subjects for training and the rest are used for the testing task. We selected nine joints from the body skeletal as shown in Figure 5.1. These joints were used as an initial features input for descriptor. The number of hidden neurons were selected by experiment to perform high accuracies and our results are compared with state- of-the- arts methods that rely only on the skeleton joints description.

5.3.1 Recognition system

We employ Extreme Learning Machine ELM for the action classification. ELM is a multi-class classifier recently introduced for pattern recognition. The proposed action recognition system incorporates this classifier, which is a version of the feed forward neural network [Huang et al., 2012]. Compared with other classifiers, ELM provides significant performances, such as fast learning time and recognition accuracy.

In [Harris and Stephens, 1988], ELM was adopted for human activity recognition from video data. In recent years, this learning algorithm has been applied to solve skeleton-based human action recognition problems [Chen and Koskela, 2013] and many other computer vision problems. In this section, we present a brief review of the theory underlying this type of machine learning. For more details about the classical materials of ELM, see [Huang et al., 2006].
We summarize the mathematical sounds of ELM as follows. When the training sample A is given by (x_j, y_j), $j = [1, \ldots, q]$, in which $x_j \in R^N$ and $y_j \in R^M$, the output function of ELM model with L hidden neurons can be expressed as follows:

$$f_l(x) = \sum_{i=1}^{L} g_i \omega_i(x) = \mathbf{\Omega}(\mathbf{x})\ \mathbf{G}. \tag{5.12}$$

where $\mathbf{G} = [\mathbf{g_1}, \ldots, \mathbf{g_L}]$ is the output weight vector relating the L hidden nodes to the $m > 1$ output nodes, and $\mathbf{\Omega}(x) = [\omega_1(x), \ldots..\omega_L(x)]$ is a nonlinear activation function. The system $\mathbf{\Omega}_i(x)$ can be written in an explicit form presented as follows:

$$\mathbf{\Omega}_i(x) = \beta(\tau_i.x + \epsilon_i), \tau_i \in R^d, \epsilon_i \in R. \tag{5.13}$$

where $\beta(.)$ is an activation function with hidden layer parameters (τ, ϵ). In the second stage of ELM learning, the error minimization between training data and output weight Ω is solved by using the least square norm depicted below.

$$min\|\mathbf{\Omega G} - \mathbf{H}\|^2, \mathbf{G} \in \mathbf{R}^{N*M}. \tag{5.14}$$

where$\mathbf{\Omega}$ defines the system of the layer of hidden neurons given as

$$\mathbf{\Omega} = \begin{bmatrix} \beta(\tau_1.x_1 + \epsilon_1) & \ldots & \beta(\tau_L.x_1 + \epsilon_L) \\ \vdots & \ddots & \vdots \\ \beta(\tau_1.x_N + \epsilon_1) & \ldots & \beta(\tau_L.x_N + \epsilon_L) \end{bmatrix}. \tag{5.15}$$

and $\mathbf{H}$ is the training data matrix denoted as

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1^{\mathbf{T}} \\ \vdots \\ \mathbf{h}_{\mathbf{N}}^{\mathbf{T}} \end{bmatrix}. \tag{5.16}$$

The optimal solution for minimizing the training error in (5.14) practically assumes that the number of hidden neurons L is less than that of the training set (i.e., $L < Q$). Therefore, in using the Moore–Penrose generalized inverse of matrix $\mathbf{\Omega}$, the optimal solution for (5.14) is given by[Huang et al., 2012]:

$$\mathbf{G}^* = \mathbf{\Omega}^*\mathbf{H}. \tag{5.17}$$

Where $\mathbf{\Omega}^*$ is the inverse of $\mathbf{\Omega}$.

5.3.2 MSR Action 3D dataset

Previous recognition results have already been reported in the literature using the MSRAction3D dataset. Table 5.1 shows the recognition rate per action subset along with the corresponding results of methods that rely on skeleton joints. As we can see, our method gives a good results. More specifically, our method outperforms most of the state-of-the-art methods on this dataset. Individually, the proposed method achieves 90.94 % which is higher than the most state-of-the-arts reported in [Xia et al., 2012, Yang and Tian, 2012, Ohn Bar and Trivedi, 2013, Hussein et al., 2013], but it is slightly lower than the recent result reported in [Vemulapalli et al., 2014]. In this case 750 hidden layers are observed in ELM The proposed method significantly improves action recognition accuracy in comparison to the accuracies of the existing methods.

Table 5.1: Comparison of Recognition rates with the state-of-the-art results on MSR action dataset

Histograms of 3D joints [Xia et al., 2012]	78.97
EigenJoints [Yang and Tian, 2012]	82.30
Joint angle similarities [Ohn Bar and Trivedi, 2013]	83.53
Covariance descriptors [Hussein et al., 2013]	90.53
Random forests [Zhu et al., 2013]	90.90
Joints as special Lie algebra [Vemulapalli et al., 2014]	92.46
Proposed approach	**90.94**

5.3.3 UTKinect Action Dataset

Similar to [Zhu et al., 2013], we experimented with our approach on a UTKinect-Action. Table 5.2 summarizes the recognition accuracies of our method compared with current skeleton-based method using UTKinect dataset. In this case the proposed approach gives the best results on these datasets. For example, the average accuracy of our method outperforms the average accuracy of [Xia et al., 2012] and [Zhu et al., 2013] by 0.73% and 3.75%, respectively. The number of the hidden layers was 700 for this dataset.

Table 5.2: Comparison of Recognition rates with the state-of-the-art results using UTKinect dataset

Random forests [Zhu et al., 2013]	87.90
Histograms of 3D joints [Xia et al., 2012]	90.92
Proposed approach	**91.65**

5.3.4 Florence Action dataset

We further evaluate our method using Florence dataset, the recognition rates compared with various methods were reported in Table 5.3, The proposed method gives the best over the results of [Seidenari et al., 2013] by 5.5%. While, an algorithm of Vemulapall et. al. [Vemulapalli et al., 2014] actually achieves much higher recognition accuracy on this complex action set. The number of hidden layers in this experiment t is 820.

Table 5.3: Comparison of Recognition rates with the state-of-the-art results, using Florence dataset

Multi-Part Bag-of-Poses [Seidenari et al., 2013]	82.00
Joints as special Lie algebra [Vemulapalli et al., 2014]	90.88
Proposed approach	87.50

5.3.5 G3D dataset

We carried The last experiment on G3D-Action dataset. The average accuracy of our representation reported in Table 5.4 is 21.26%. This result is better than the average accuracy of [Bloom et al., 2012]. These results clearly demonstrate the performance of our proposed method over a number of existing skeletal joints-base approaches.

Table 5.4: Comparison of Recognition rates with the state-of-the-art results, using G3D dataset

Hybrid joints feature + adaboost [Bloom et al., 2012]	71.04
AL alwani et. al. [Alwani et al., 2014]	80.55
Proposed approach	**92.30**

5.4 Conclusion

From the experimental results observe we observe that, the Covariance descriptor on SHs typically works better than most of the existing methods. This confirms that relations between individual joint's features and harmonics motion of these features are informative and useful for action recognition. The combination of the covariance with SHs improves action recognition accuracy. This confirms that the proposed SHs directly models temporal features and the covariance descriptor models relations between features. Moreover, the use of the SHs is very important for modeling the angular orientations of the skeleton joint along temporal variation.

The problem of skeleton body representation was explicitly modeled in this paper. We have presented an efficient approach for skeleton-based human action recognition. We adopted the spherical harmonics and covariance technique. We used the spatiotemporal spherical harmonics that characterize the spherical angles of local joints over the entire action sequence. We exploited the idea of covariance components in order to capture the dynamic of the action and provide a relevant descriptor with the a fixed length.

The experimental results tested on a various datasets prove the effectiveness of the proposed method. Results demonstrate that our method can be successfully used for capturing temporal changes in action and achieve a higher recognition rate. In future studies, we will enhance our method for classifying and recognizing different other behaviors.

Chapter 6

Spatio-temporal representation of 3-D skeleton joints-based action recognition using modified spherical harmonics

Contents

In this study, we present a novel skeleton joint-based representation of 3D human action in a spatiotemporal manner. We employ the spherical angles of body joints computed from the 3D coordinates of skeleton joints. The proposed feature representation is a combination of the modified spherical harmonics (MSHs) and the spatiotemporal model of sequence level. To estimate the human pose, the SHs of spherical angles provide a distinctive feature description. As such, the problem of skeleton joint representation is addressed in a spatiotemporal approach using MSHs. The proposed model simply incorporates two mechanisms to efficiently capture the temporal dynamic of joints, namely, the application of MSHs in the computed spherical angles of each pose and the construction of MSHs in a hierarchical scheme. MSHs are computed in multi-level, in which each level encodes the time window of an action sequence.

In the proposed representation of 3D human action, the selected MSHs are adopted to characterize the features in multi-levels and capture the harmonic frequency of function in a two-sphere space. Given this condition, the defined spherical angle vector of the selected joints may be projected onto S^2. However, the principle computation required in this space is extremely large because each selected joint is sampled by the feature vectors of $MJ = \{M_1, \ldots, M_K\}, M \in R^{N \times N}$; where M, MSHs matrices of k levels, J joints numbers, and N numbers of farms in each level. Considering that the desired descriptor dimensionality aims to expedite the classification phase as well as reduce the noise and redundant feature sizes, we apply dynamic time wrapping (DTW) to determine the optimal alignment between the sublevels of hierarchical MSHs.

An action classification is performed using the extreme learning machine (ELM) classifier. The proposed method is evaluated based on recent skeleton-based 3D action datasets.

6.1 Introduction

In the previous, we present the motivation behind using covariance on SHs for action recognition. The method is mainly focused on the temporal property of local joint only to extract the skeletal feature. We also use the classical covariance to measure the relation between individual joint. However, the captured SHs along temporal variation may not be enough to capture sufficient information in a complex motion which require the fusion of the spatial distribution with temporal dynamics. The information in spatiotemporal domain might carry complementary information to each other.

Spatiotemporal representation of an action sequence can be seen as an extension of the spatial domain to incorporate temporal dimension. It measure all kinds of possible relationships between features. We introduce a new local spatiotemporal descriptor for skeleton joint, and we propose a new approach for action recognition based on the introduced descriptor. The descriptor is based on a modified

SHs basis function. which model the harmonics function in quadratics basis function.

The proposed descriptor can be used to represent skeleton joint orientations and displacement features. In order to addresses the structural information of the skeleton sequence, we use spatiotemporal domain, and we compute a Modified MSHs on joint orientations. Similar to the previous work, spherical angles (or orientations is the same) are estimated for local and global body joints, and the spatiotemporal system of these orientation is built. Then the MSHs is applied on this system. Moreover, we encode the temporal variation and different frame sequence length by construct the MSHs in hierarchical fashion. The main difference bertween the work in chapter 5 and the present one is that we use the proposed MSHs with hierarchical fashion in spatiotemporal dimension.

We present an evaluation of our approach on four various state-of-the-art datasets. We present that the MSHs achieves better than the previous SHS-based work and the the state-of-the art algorithms.

6.2 Proposed approach

6.2.1 Spatiotemporal system of joint level features

In this section, we present the extraction of joint level features in spatiotemporal domain. As mentioned in chapter 5 the skeleton joints are firstly represented in terms of the spherical angles relatively measured with respect to the fixed coordinates, which are more accurate than the joint coordinates or joint differences. The spherical angles are quantified in the spherical coordinate as illustrated in equation 5.8. All angles are computed corresponding to the origin reference (i.e., the origin of the spherical coordinate system is placed at the $hip - center$ joint coordinate). Only a primitive set of the supported joints is used for the 3D pose representation as labeled in the right side of Figure 5.1.

To further analyze the 3D skeleton joints in terms of their spatiotemporal domain, we construct a spatiotemporal system which incorporates static and dynamic movements of the body skeleton joints.

Assume that the spherical angels are available in each frame. Let the entire skeleton body be represented by J joints (i.e., $J = (1, 2, ..., K)$, and the action be performed over T frames. Thus, the spherical angle system of the entire action sequence can be constructed as a spatiotemporal system expressed as

$$F_{s\in A}(\theta,\phi) = Pose\Bigg\downarrow \begin{array}{c} J_1 \\ J_2 \\ \vdots \\ J_k \end{array} \begin{bmatrix} (\theta,\phi)_{1,1} & (\theta,\phi)_{1,2} & \dots & (\theta,\phi)_{1,T} \\ \theta,\phi)_{2,1} & (\theta,\phi)_{2,2} & \dots & (\theta,\phi)_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ (\theta,\phi)_{J,1} & (\theta,\phi)_{J,2} & \dots & (\theta,\phi)_{J,T} \end{bmatrix}. \qquad (6.1)$$

where s is the specific action, T is the total number of frames in the action sequence, and J is the total number of joints in the static pose or frame. In the above equation, each row represents the spherical angles of the local joint displacement in the time sequence, while each column depicts the spherical angles of each pose in the action sequence.

The representation based on the above system features provides a rotation invariant representation of an action sequence. However, the relationships between these joint level features and the spatial positions of these features may be informative and useful for action recognition.

6.2.2 Modified SHs

As mentioned, this study proposes a novel feature extraction framework, in which the modified real part notation of SHs is used to represent the spatiotemporal features of skeleton joints and improve human action recognition. However the term real function of Standard SHs is given as:

$$y_n^m(\theta,\phi) = \sqrt{(2)}Q_n^m \cos(m\phi) Z_n^m(\cos\theta). \qquad (6.2)$$

Where Q_n^m is the scaling factors expressed as:

$$Q_n^m = \sqrt{\frac{(2n+1)(n-|m|)!}{4m(n+|m|)!}}. \qquad (6.3)$$

The real part function $\cos(m\phi)$ of SHs, may be expanded using the trigonometric identity into the following expression

$$\cos(2\phi) = 2\cos^2\phi - 1. \qquad (6.4)$$

Put 6.4 in 6.2, the modified SHs has the following form:

$$y_n^m(\theta,\phi) = Q_n^m[2\cos^2\phi - 1]Z_n^m\cos\theta. \tag{6.5}$$

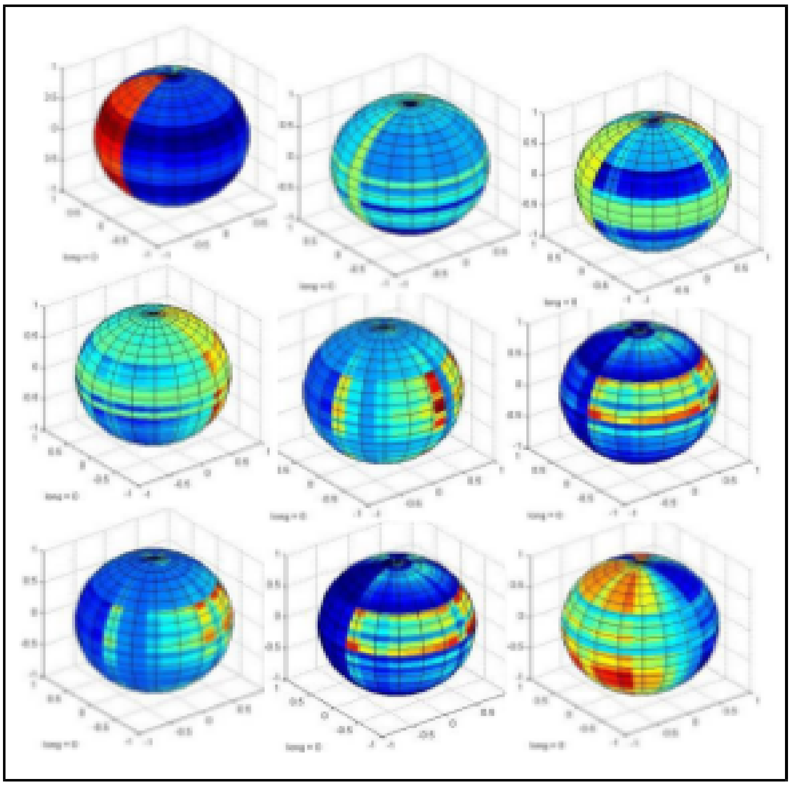

Figure 6.1: Examples of harmonics basis function for a person performs a tennis swing action. (Top panel/ left to right) temporal representation of :Elbow Right/Left, Wrist Right.(Middel panel), wrist Left, Knee Right/left. (Bottom panel), Foot R/L , and Head Joints respectively

Where Q is the scale factor, and Z is the associated Legendre polynomials. The quadratic term in 6.5 captures the angular velocity of joint displacement. This velocity is useful to differentiate the actions involved in a curved motion, such as waving or shape drawing. Thus, for a given action, the angular quantities (e.g., relative angular speed and changes in directions of these joints) can be more stable across objects than their actual 3D positions.

However, the MSHs of the local 3D skeleton joints capture discriminant information about different actions. In other words, the quadratic term in MSHs describes the direction and angular speed of joint motions. Experiments have proven that introducing the quadratic angular velocity and direction of joint dynamics significantly improves the use of the standard SHs.

For the system depicted in 6.1, we compute its MSHs basis functions as explained in equations 5.8-5.10, with exception that in equation 5.10, instead of using the standard form, we use equation 6.5 for $m = 2$.

The estimated MSHs for the body pose at time t (each column of 6.1) form the static pose features descriptor. The collection of the estimated MSHs for all frames of a specific action defines the static poses representation vectors of $\mathbf{H^s} = [\mathbf{P_1}, \mathbf{P_2}, ..., \mathbf{P_T}]$. Similarly, the MSHs of the local joint displacement are calculated by projecting each row of equation 6.1 onto the basis functions of MSHs. In this case, the individual MSHs of each local joint displacement is calculated over the entire row of equation 6.1. To form the MSHs vector of the local joint motion for a given action segment, we collect the individual motion vectors $\mathbf{H^m} = [\mathbf{M_1}, \ldots, \mathbf{M_J}]$.
Figure 6.1 shows the real example of the MSHs calculation on the individual joint for subject which performs tennis action. In this figure, each sphere demonstrates the harmonics distribution corresponding to the individual joint listed in equation 6.1. We can see from Figure 6.1 the ability of MSHs to discriminate the temporal variations between local joints.

6.2.3 Temporal construction of MSHs in hierarchical model

In 3D skeleton-based action recognition, a compact skeleton-based descriptor should encode the static pose information and the temporal evolution or joint motion at a given time segment. The static pose and joint displacement features of a given skeleton body sequence contain discriminative data about the human action over a time segment.
In the previous section, the MSHs capture the spatial dependency of the holistic joints (i.e., pose in frame) and the motion of the local joint properties over the time sequence. To efficiently encode the temporal variation of the local joints over time, each SH of these joints is constructed in a hierarchical manner. The idea of hierarchical construction is inspired by the spatial pyramid matching introduced by [Lazebnik et al., 2006] to achieve matching in 2D images. Relying on determining the MSHs calculated in the previous section, we construct the MSHs of the local joints in a multi-level approach. Each MSHs covers a specific time window of the action sequence. The MSHs are computed over the entire video sequence from the top level and over the smaller windows at the lower levels. Window overlapping is used to increase the ability of the proposed representation to dif-

ferentiate multiple actions by sliding from one window to the half of the next one, as depicted in Figure 6.2.

Regardless of whether the multiple levels of SHs are used, differentiating the local temporal sequences of various action categories is a difficult task because of numerous issues, including the frame rate variations and the temporal independence in each sub-level. To address these issues, DTW [Muller,] is used to compute for a distance between the multiple levels of SHs for each action category. Similarly, DTW is used to identify the nominal distances between the SHs of consecutive levels for each local joint. The distance vector for each local joint displacement is then formed. The temporal model of the skeleton joints is encoded for each action category as a concatenation of the distance vector $\mathbf{D^t} = [\mathbf{T_1}, \dots, \mathbf{T_J}]$. Through the computation of the pose and motion feature vectors of the whole skeleton joints, an action sequence is represented by a combination of these vectors to form a skeleton representation features vector as

$$\mathbf{S} = \mathbf{H^s} + \mathbf{D^t}. \tag{6.6}$$

The static pose and temporal dynamic of the harmonics contain information about the spatiotemporal function over a time sequence of an action. Therefore, this type of harmonic information can be considered as a compact representation of the body skeleton joint and can be used to reliably classify an actions.

6.2.4 Alternative body skeleton features

Alternative skeleton representations are adopted as an another abstraction of the skeleton features which are used for further performance evaluation of our method. These skeleton representations are as follows:
Joint Location (JL): simply concatenates all joint coordinates in one vector.
Pairwise joints differences (PJDs): concatenation $y_f = \{p_i - p_j | i, j = (1, 2, \dots, K), i \neq j\}$ of all frames.
Magnitude of the Position Velocity (MPV): the velocity between the same joints of enter frame defined as $Y_{t1,t2} = ||p_{i,t1} - p_{i,t2}||$.
These skeleton representations are fed directly into the classifier to directly compare the proposed method with the alternative representation schemes.

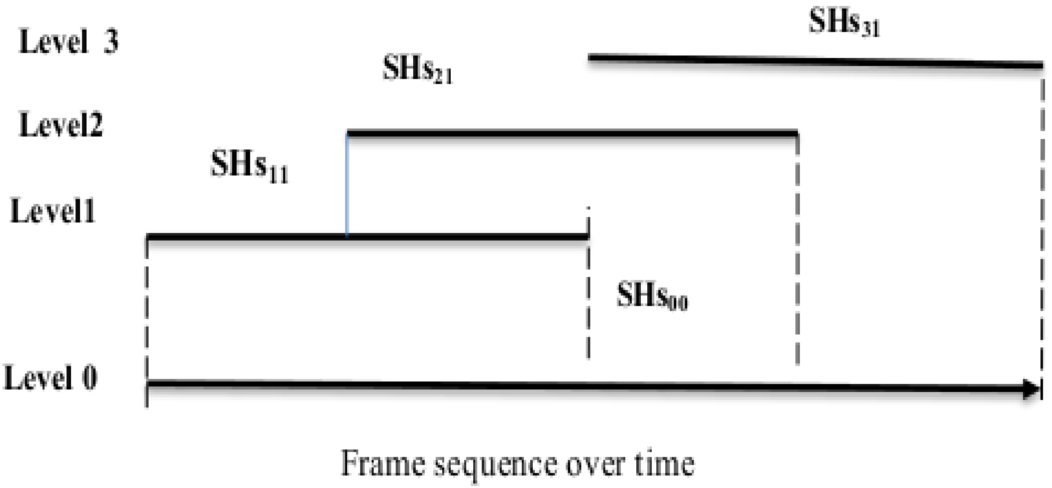

Figure 6.2: A 3-level representation of Temporal construction of the SHs, SHs$_{l}j$ is the jth Spherical harmonics in the lth level of the hierarchy

6.3 Experimental results

To evaluate the effectiveness of the proposed method, we perform action recognition on the proposed feature representation and recently published datasets (i.e., MSR-Action 3D, G3D, Florence 3D Action, and UTKinect-Action datasets). These datasets are used as benchmarks in the experiment. The action complexity of these datasets varies from simple to complex sequences. In addition to depth data, skeleton data are also provided by these datasets using a Kinect sensor as required.

In all experiments, an ELM classifier is used with the proposed representation. For each dataset, the state-of-the-art skeleton-based methods are extensively compared with the proposed approach. The number of hidden neurons of ELM is experimentally tuned for each dataset. To simplify the computation in each experiment, we set the frequency of the basis functions over the sphere equal to $n = 2$ and the degree $m = 2$. We consider the cross subject protocol for the test setting in all datasets. In particular, half of the subjects are used for training, and the other half for testing. In all experiments, we use nine joints from the body skeletal as the initial input joints to our proposed method, as shown in Figure 6.3.

The features from these joints are initially used for the skeleton feature representation.

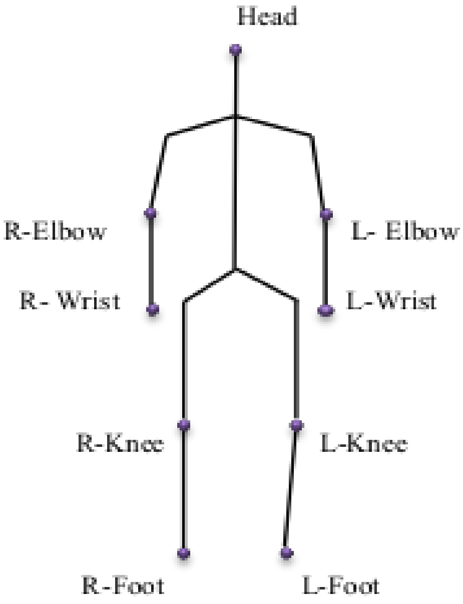

Figure 6.3: Labeled skeleton joints used as initial input to the proposed method

Table 6.1: Recognition rates for various skeletal representations on MSR-Action3D dataset

subset	JL	PDJ	MPV	Proposed
AS1	72.2	76.22	80.23	**89.76**
AS2	69.83	80.47	79.15	**91.7**
AS3	82.7	71.4	84.06	**92.5**
Average	74.91	72.36	81.14	**90.98**

Table 6.2: Recognition rates for various skeletal representations on UTKinect Action, Florence3D Action, and G3D Action datasets

Dataset	JL	PDJ	MPV	Proposed
UTKinect	82.5	83.08	87.58	**93.0**
Florence3D	76.59	70.33	83.7	**86.13**
G3D	79	80.36	82.04	**92.89**

Table 6.3: Comparison of Recognition rates with the state-of-the-art results on MSR-Action3D dataset

Approaches	**Accuracy**
Xia et. al 2012 [Xia et al., 2012]	78.97
yang & Tian 2012 [Yang and Tian, 2012]	82.30
Ohn & Trivedi 2013 [Ohn Bar and Trivedi, 2013]	83.53
Zhu et. al 2013[Zhu et al., 2013]	90.90
Hussien et. al.2013 [Hussein et al., 2013]	90.53
Evangelidis at. al 2012 [Evangelidis et al., 2014]	89.86
Vemulapali et. al 2014 [Vemulapalli et al., 2014]	92.46
SHs [AL alwani and Chahir 2015] [Alwani and Chahir, 2015]	90.94
proposed approach	**90.98**

6.3.1 Comparison with various skeleton features

The performance of various representations is evaluated on all datasets, and the efficiency of the proposed method is compared with that of other skeleton representations. Table 6.1 reports the accuracy of the proposed approach with the corresponding results of different representation methods based on the MSR-Action dataset. Our findings presented in this table are achieved using three levels of SHs, while the window overlap in the second and third levels is preserved. Compared with other skeleton representations, the proposed method provides satisfactory results. In particular, the proposed method improves the average accuracies of JL, PDJs, and MPV by 16.07%, 18.62%, and 9.84% , respectively. These observations clearly indicate the superiority of the proposed representation over existing skeleton representations.

Tables 6.2 summarizes the recognition accuracies of various skeleton representations on the UTKinect-Action, Florence 3D Action, and G3D datasets. The results reveal that our method significantly outperforms the other skeleton representations on these datasets. In using UTKinect dataset, the accuracy of the proposed representation is 10.5% better than that of JL, 9.92% better than that of PJDs, and 5.42% better than that of MPV. In the case of the Florence dataset, the accuracy of the proposed representation is 9.54%, 15.8%, and 2.43% better than that of JL, PJDs, and MPV, respectively. In the case of the G3D dataset, the accuracy of the proposed representation is 13.83% better than that of JL, 12.47% better than that of PJDs, and 10.79% better than that of MPV.

6.3.2 Comparison with the state-of-the-art

The same datasets are used to compare the performance of the proposed method with those of existing state-of-the-art methods. For each data set, the hidden neurons are reported separately. In all experiments, the results correspond to using three levels of hierarchical SHs, while preserving the overlap in the last two levels.

Several recognition results on the MSR-Action 3D dataset are already available in the literature. Table 6.3 presents the recognition rate of the proposed approach along with those of the corresponding current methods. As indicated in this table, the proposed approach obtains the best results compared with those of most existing methods. In particular, our method provides good results in line with those of some existing methods but outperforms the others. In this case, 780 hidden neurons are observed in ELM.

For further evaluation, the proposed approach is applied to the skeleton sequences from UTKinect-Action, Florence, and G3D Action datasets. The performance of the proposed approach in this experiment is also compared with those of the corresponding methods. Table 6.4 compares our method with various state-of-the-art skeleton-based human action recognition approaches on the*UTKinect dataset*. The proposed approach gives comparable results. The average accuracy of the proposed representation is 5.10% better than that given in [Zhu et al., 2013] and 2.08% better than that in [Xia et al., 2012]. The number of hidden neurons in this experiment is 640.

Table 6.5 reports the average recognition accuracies in the case of the Florence dataset. The results reveal that the accuracy of the proposed method is slightly higher than that cited in [Seidenari et al., 2013]. In particular, the performance of the proposed approach is superior over that of the state-of-the-art methods by 4.13%. Our results in this table correspond to 500 hidden neurons for ELM.

The performance of the proposed method is also assessed based on the G3D-Action dataset. Table 6.6 demonstrates the results, which indicate that our method evidently outperforms the existing skeletal joint-based state-of-the-art methods by achieving better accuracy by 0.59%. In this experiment, 700 hidden neurons exist in the ELM.

Table 6.4: Comparison of recognition rates with the state-of-the-art results using UTKinect dataset

Zhu et. al 2013 [Shimada and Taniguchi, 2008]	87.90
Xia et. al 2012 [Xia et al., 2012]	90.92
Devianne et. al 2013 [Devanne et al., 2013]	91.5
SHs [AL alwani and Chahir 2015] [Alwani and Chahir, 2015]	91.65
Proposed approach	**93.0**

Table 6.5: Comparison of recognition rates with the state-of-the-art results, using Florence dataset

Siednari et. al 2013 [Seidenari et al., 2013]	82.00
SHs [AL alwani and Chahir 2015] [Alwani and Chahir, 2015]	87.50
Proposed approach	86.13

Table 6.6: Comparison of recognition rates with the state-of-the-art results, using G3D dataset

Bloom et. al 2012 [Bloom et al., 2012]	71.04
AL alwani et. al 2014 [Alwani et al., 2014]	80.55
SHs [al alwani and Chahir 2015] [Alwani and Chahir, 2015]	92.30
Proposed approach	**92.89**

6.3.3 Benefit of modified SHs

Table 6.7 demonstrates that the addition of dynamic features expressed by the second-order term of the real SHs dramatically increases the recognition accuracy compared with the standard SHs [Alwani and Chahir, 2015]. The efficiency of using MSHs becomes evident when we compare them with the standard SH descriptors. In Table 6.4, the recognition accuracies of MSHs are used and compared with those of the standard SHs. The explicit estimation of angular speed and directions in terms of the second-order function presents a significant performance. For example, in the MSR-Action 3D dataset, the use of the quadratic term in MSHs improves the recognition accuracy by a substantial .04% margin over the standard SHs. In the case of the UTKinect and G3D datasets, the MSHs add a significant improvement of 1.35% and 0.59% to their recognition accuracies respectively. Contrarily, in the Florence dataset, the recognition rate is decreased from 87.5% for SHs to 86.13% for MSHs.

Our findings affirm that the angular speed component of the quadratic function is extremely important for action representation with curved displacement. Such a displacement cannot be fitted by the spatiotemporal features of the standard real SHs.

Table 6.7: Comparison of Recognition rates with the SHs-based state-of-the-art results

Datasets	**SHs**	**MSHs**
MSR Action 3D	90.94	**91.98**
UTKnect	91.65	**93.00**
Florence	**87.50**	86,13
G3D	92.30	**92.89**

6.4 Conclusion

In this chapter, we introduced a novel framework for action recognition based on an explicit model of 3D skeleton joints in a spatiotemporal domain. SH transform was used to explicitly model the angular speed and direction of joints. A novel MSH was also proposed based on the quadratic function of the real part of SHs. According to our framework, all body joints were registered into a body coordinate system to extract the spherical angles of joints expressed in 3D body coordinates. The spatiotemporal system of human action was then constructed and encoded by a set of MSHs of static poses and local joint displacement over time. The temporal evolution of the skeleton joints was characterized in a hierarchical manner. An appropriate skeleton representation of an action was formulated as a vector of combined poses and joint motion features. For the action recognition, ELM was used. The performance of the proposed method was evaluated based on recent 3D skeleton-based datasets. We compared the proposed method with the existing state-of-the-art methods either by adopting pure skeleton data or by directly relying on depth data. The experimental results revealed that depending on the used datasets, the proposed method can obtain results similar to those of the extant methods or outperform them. The findings also indicated that MSHs are well suited for action representation with curved movement and angular direction changes.

In summary, our newly proposed method is in line with the recently presented methods for 3D skeleton-based pose representation. The angular direction estimated from skeleton data and its derived SHs are relevant for action recognition and can be successfully used to capture temporal changes in action and obtain a high recognition rate.

CHAPTER 7

CONCLUSION AND PERSPECTIVES

Contents

7.1 Summary of our contributions

The central motivation of the thesis is human action and event recognition. We have addressed this problem from the perspective of features representations for both thermal and 3D RGBD imaging, and we have proposed temporal-based feature encoding methods for event recognition in thermal video. We have proposed three skeleton-based features representations algorithms for human action recognition in RGBD video.

• **Temporal-based for analyzing thermal images over time** : The first challenge addressed was the real event recognition in Neonatal Intensive Care system NIC from thermal video. We have introduced two feature descriptors based on local temporal evolution of thermal signature for event recognition. The first one is based on the non redundant local binary pattern. Based on the fact that facial region and temperature changes features are the main cues of an even, we explicitly extract these features from thermal video sequences by NRTLBP. In order to quantize NRLBP, we choose the maximum and minimum channels of local temperature values as initial raw thermal input to the NRTLPB descriptor. Then, we have extended the idea of NRTLBP from the time domain to the wavelet domain, and proposed a wavelet NRTLBP.

An event in NIC-based thermal video is viewed as a temporal variation patterns in temporal dimensions. To effectively capture the temporal information of event

in NIC, we have further proposed topological persistence of 1D technique. To this end, the proposed method able to extract useful information from a large set of thermal noisy features. These approaches were more applicable to real event recognition tasks. The proposed methods were also shown to be compact to encode different type of thermal measurement, and to offer considerably improved performance on challenging thermal scene benchmarks.
We have presented a performance evaluation of the above techniques, and we have demonstrated that the proposed methods obtain better accuracy on real thermal-based NIC dataset.

• **Skeleton-based Human action recognition from RGBD video** : The next challenge was the investigation a novel problems of recognizing human actions from a body skeleton joints using RGBD data. We first proposed skeleton joint-based 3D action recognition framework. We developed 3D body reference coordinates system by projecting the real world coordinates into skeleton space. The set of the primitive joints are selected and the angels between these joints are computed in orthogonal planes, the angls are then concatenated into a feature vector. This feature is used as an abstraction of the body skeleton joint.
Since joint positions or distances between them, are not always provide good joint representation for complex actions, we Designed two explicit approaches for skeleton-based human action recognition. The first approach describes the temporal evolution of local joint using spherical harmonics basis functions SHs. Interesting spherical orientations of local joint are estimated in temporal domain and described using spherical harmonics basis. Furthermore, to effectively capture the dependency between joints, we have proposed covariance descriptor on SHs for the final representation of skeleton-based actions.
We have presented a performance evaluation of the above approach, and we have shown that the proposed methods obtain better or similar performance in comparison to the existing state-of-the-arts methods on various 3D action datasets.

Our last key contribution of skeleton-based human action recognition consists of modified SHs approach to encode human actions in spatiotemporal domain. To accomplish this, we have developed the spherical orientation of the selected joints as a spatiotemporal system. Then we have introduced MSHs in a hierarchical mode to cope the temporal variation, noise, and frame length variability. Our experiments have shown that this approach outperforms the current state-of-the-art methods. From the obtained results, we can conclude that spatiotemporal relations and harmonics basis bring a significant improvement over the alternative joint representation. In addition, we have shown that formulating the skeleton-based recognition problem as an explicit model problem allows to take into account any relationships between local features (e.g. spatiotemporal and/or spatial relationships).

7.2 Limitations

The main limitations of the event recognition in thermal video approach is the requirements of the automatic facial segmentation and tracking, robust 1D signal segmentation, and fusion multiple physiological behavior responses. These limitations may be possible on challenging datasets such Pretherm dataset.

The main limitations of the proposed 3D skeleton joint features description are lack of the precision, body part occlusion, and a low accuracy of joint position tracking in more complex scenario.

7.3 Perspectives

In term of perspectives, we feel it important to investigate :

- **Multimodal imaging in medicine** : The human body is homeothermic, i.e. self-generating and regulating the essential levels of temperature for survival. Thermal imaging offers the great advantage of real time two-dimensional temperature measurement. The credibility and acceptance of thermal imaging in medicine is subject to critical use of the technology and proper understanding of thermal physiology. A representative data set of large group needs to be collected and tested for evolving medical applications for thermal imaging, including inflammatory diseases, complex regional pain syndrome.
- **Evaluation of skeleton-based approaches** . We would like to evaluate our approaches on other challenging datasets, such as (MHAD) [Ofli et al., 2013], HDM05-MoCap Dataset [Müller et al., 2007], and MSRC-12 Kinect Gesture Dataset [Fothergill et al., 2012].
- **Towards automatic prediction of action segment** . The possible direction

including: developing an automatic segmentation method for actions, so that when the actor is performing continuously, we will be able to detect the beginnings and ends of the actions. In addition, instead of running the recognizer after the whole action has been performed, we will extend the system to predicting the actions during performance, which will provide further valuable information in on line applications.

- **Action modeling with multiples cues** . In chapters 4, 5, and 6, we discussed recognition performance of the proposed methods and concluded that each method has specific characteristics that could benefit from an adapted description. This has been especially obvious for spherical harmonics-based techniques. Consequently, it seems necessary to adapt multiples actions representation. One aspect is the combination of skeletal with other cues, such as silhouette, body structure, and motion.

- **Relative Trajectories of body joints in dynamic coordinate systems** . The possible direction is to investigate the Relative Trajectories using various dynamic coordinate systems (e.g. human body center), and using several dynamic coordinate systems at the same time, what could additionally enhance the discriminative power of trajectories.

Bibliography

[Aggarwal and Ryoo, 2011] Aggarwal, J. and Ryoo, M. (2011). Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43.

[Ahonen et al., 2006] Ahonen, T., Hadid, A., and Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041.

[Al Alwani et al., 2014] Al Alwani, A., Chahir, Y., and Jouen, F. (2014). Thermal signature using non-redundant temporal local binary-based features. In *ICIAR14*, pages II: 151–158.

[Ali et al., 2007] Ali, S., Basharat, A., and Shah, M. (2007). Chaotic invariants for human action recognition. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8.

[Alwani and Chahir, 2015] Alwani, A. A. and Chahir, Y. (2015). 3-d skeleton joints-based action recognition using covariance descriptors on discrete spherical harmonics transform. In *ICIP2015*.

[Alwani et al., 2014] Alwani, A. A., Chahir, Y., Goumidi, D. E., Molina, M., and Jouen, F. (2014). 3d-posture recognition using joint angle representation. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems - 15th International Conference, IPMU 2014, Montpellier, France, July 15-19, 2014, Proceedings, Part II*, pages 106–115.

[Andreone et al., 2002] Andreone, L., Antonello, P., Bertozzi, M., Broggi, A., Fascioli, A., and Ranzato, D. (2002). Vehicle detection and localization in infrared images. In *Proc. IEEE International Conference on Intelligent Transportation Systems*, page 141–146.

[Andres et al., 2013] Andres, S., Conrad, S., Mehrtash, T. H., and Brian, C. L. (2013). Spatio-temporal covariance descriptors for action and gesture recognition. *CoRR*, abs/1303.6021.

[Antônio et al., 2012] Antônio, W. V., Thomas, L., William, S., and Mario, F. M. C. (2012). Distance matrices as invariant features for classifying mocap data. In *ICPR 2012 (21st International Conference on Pattern Recognition)*, pages 2934–2937. IEEE.

[Ballin et al., 2012] Ballin, G., Munaro, M., and Menegatti, E. (2012). Human action recognition from rgb-d frames based on real-time 3d optical flow estimation. In Chella, A., Pirrone, R., Sorbello, R., and Johannsdottir, K. R., editors, *BICA*, volume 196 of *Advances in Intelligent Systems and Computing*, pages 65–74. Springer.

[Barnachon et al., 2013] Barnachon, M., Bouakaz, S., Boufama, B., and Guillou, E. (2013). A real-time system for motion retrieval and interpretation. *Pattern Recognition Letters, special issue on Smart Approaches for Human Action Recognition*.

[Bertozzi et al., 2003] Bertozzi, M., Broggi, A., Grisleri, P., Graf, T., and Meinecke, M. (2003). Pedstrain detection in infrared images. In *IEEE Intelligent Vehicles Symp., Columbus*.

[Bingbing et al., 2013] Bingbing, N., Gang, W., and Pierre, M. (2013). Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *In Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer.

[Blank et al., 2005] Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402.

[Bloom et al., 2012] Bloom, V., Makris, D., and Argyriou, V. (2012). G3d: A gaming action dataset and real time action recognition evaluation framework. In *CVPR Workshops*, pages 7–12. IEEE.

[Bobick and Davis, 2001] Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):257–267.

[Brechbühler et al., 1995] Brechbühler, C., Gerig, G., and Kübler, O. (1995). Parametrization of closed surfaces for 3-d shape description. *Computer Vision and Image Understanding*, 61(2):154 – 170.

[Bustos et al., 2005] Bustos, B., Keim, D. A., Saupe, D., Schreck, T., and Vranić, D. V. (2005). Feature-based similarity search in 3d object databases. *ACM Computing Surveys*, 37(4):345–387.

[Chaquet et al., 2013] Chaquet, J. M., Carmona, E. J., and Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659.

[Chaudhry et al., 2013] Chaudhry, R., Ofli, F., Kurillo, G., Bajcsy, R., and Vidal, R. (2013). Bio-inspired dynamic 3d discriminative skeletal features for human action recognition. pages 471–478.

[Chen and Koskela, 2013] Chen, X. and Koskela, M. (2013)). Skeleton-based action recognition with extreme learning machines. In *International Conference on Extreme Learning Machines*.

[Dai et al., 2005] Dai, C., Zheng, Y., and Li, X. (2005). X.: Layered representation for pedestrian detection and tracking in infrared imagery. In *In: IEEE CVPR WS on OTCBVS (2005*.

[Dai et al., 2001] Dai, Y., Shibata, Y., Ishii, T., Hashimoto, K., Katamachi, K., Noguchi, K., Kakizaki, N., and Cai, D. (2001). An associate memory model of facial expressions and its application in facial expression recognition of patients on bed. In *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo, ICME 2001, August 22-25, 2001, Tokyo, Japan*, page 72—75.

[Dalal et al., 2006] Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, pages 428–441.

[Davis and Keck, 2005] Davis, J. W. and Keck, M. A. (2005). A two-stage template approach to person detection in thermal imagery. In *WACV/MOTION*, pages 364–369. IEEE Computer Society.

[Devanne et al., 2013] Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., and Del Bimbo, A. (2013). Space-time Pose Representation for 3D Human Action Recognition. In *ICIAP Workshop on Social Behaviour Analysis*, page 1.

[Dollár et al., 2005] Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, pages 65–72.

[Edelsbrunner and Harer, 2008] Edelsbrunner, H. and Harer, J. (2008). *Persistent homology – a survey*, volume 453.

[Edelsbrunner et al., 2000] Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2000). Topological persistence and simplification. pages 454–.

[Evangelidis et al., 2014] Evangelidis, G., Singh, G., and Horaud, R. (2014). Skeletal quads: Human action recognition using joint quadruples. In *22nd International Conference on Pattern Recognition (ICPR)*, volume 42, pages 513–529.

[Fanello et al., 2013] Fanello, S. R., Gori, I., Metta, G., and Odone, F. (2013). Keep it simple and sparse: Real-time action recognition. *J. Mach. Learning Research*, 14(1):2617–2640.

[Farah et al., 2011] Farah, A.-K., Reza, S., Heather, E., and Derek, B. (2011). An evaluation of thermal imaging based respiration rate monitoring in children. *American Journal of Engineering and Applied Sciences*, 4(4):586–597.

[Fothergill et al., 2012] Fothergill, S., Mentis, H., Kohli, P., and Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1737–1746. ACM.

[Geert et al., 2008] Geert, W., Tinne, T., and Luc, V. G. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In Forsyth, D. A., Torr, P. H. S., and Zisserman, A., editors, *ECCV (2)*, volume 5303 of *Lecture Notes in Computer Science*, pages 650–663. Springer.

[Gorelick et al., 2007] Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29:2247–2253.

[Grazia et al., 2005] Grazia, M., Bono, D., Pieri, G., and Salvetti, O. (2005). Multimedia target tracking through feature detection and database retrieval. In *in Proceeding of the 22nd International Conference on Machine Learning-Workshop on Machine Learning Techniques for Processing Multimedia Content(ICML)*, pages 19–22.

[Gunaratne and Sato, 2003] Gunaratne, P. and Sato, Y. (2003). Estimation of asymmetry in facial actions for the analysis of motion dysfunction due to paralysis. 3(4):639–652.

[Harris and Stephens, 1988] Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151.

[Hatcher, 2002] Hatcher, A. (2002). *Algebraic Topology*. Cambridge Univ. Press.

[Huang et al., 2012] Huang, G.-B., Zhou, H., Ding, X., and Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 42(2):513–529.

[Huang et al., 2006] Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1–3):489 – 501.

[Hussein et al., 2013] Hussein, M. E., Torki, M., Gowayyed, M. A., and El-Saban, M. (2013). Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2466–2472. AAAI Press.

[Jiang et al., 2013] Jiang, Y.-G., Bhattacharya, S., Chang, S.-F., and Shah, M. (2013). High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval (IJMIR)*, 2(2):2:73–101.

[Jungling and Arens, 2009] Jungling, K. and Arens, M. (2009). Feature based person detection beyond the visible spectrum. In *IEEE CVPR Workshpos*.

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

[Lan et al., 2010] Lan, T., Wang, Y., Yang, W., and Mori, G. (2010). Beyond actions: Discriminative models for contextual group activities. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 1216–1224. Curran Associates, Inc.

[Laptev,] Laptev, I. On space-time interest points. *International Journal of Computer Vision*, 64(2/3).

[Laptev and Lindeberg, 2003] Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In *IN ICCV*, pages 432–439.

[Lazebnik et al., 2006] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, page 2169–2178.

[Lebedev N., 1972] Lebedev N., N. (1972). *Special Functions and Their Applications*. Dover Books on Mathematics.

[Li and Gong, 2010] Li, J. and Gong, W. (2010). Real time pedestrian tracking using thermal infrared imagery. *JOURNAL OF COMPUTERS*, 5(10):1606–1613.

[Li et al., 2012] Li, W., Zheng, D., Zhao, T., and Yang, M. (2012). An effective approach to pedestrian detection in thermal imagery. In *ICNC*, pages 325–329. IEEE.

[Lui and Beveridge, 2011] Lui, Y. M. and Beveridge, J. R. (2011). Tangent bundle for human action recognition. In *FG*, pages 97–102. IEEE.

[Lv and Nevatia, 2006a] Lv, F. and Nevatia, R. (2006a). Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In Leonardis, A., Bischof, H., and Pinz, A., editors, *ECCV (4)*, volume 3954 of *Lecture Notes in Computer Science*, pages 359–372. Springer.

[Lv and Nevatia, 2006b] Lv, F. and Nevatia, R. (2006b). Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV*, ECCV'06, pages 359–372. Springer-Verlag.

[Mallat, 2008] Mallat, S. (2008). *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition.

[Mark et al., 2005] Mark, A., James, K., and Davis, W. (2005). A two-stage template approach to person detection in thermal imagery. In *Proc. Wkshp. Application of Computer Vision*.

[Milnor, 1973] Milnor, J. (1973). *Morse Theory*. Princeton University Press.

[Minhas et al., 2010] Minhas, R., Baradarani, A., Seifzadeh, S., and Jonathan, W. Q. (2010). Human action recognition using extreme learning machine based on visual vocabularies. *Neurocomputing*, 73(10):1906–1917.

[Müller et al., 2007] Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., and Weber, A. (2007). Documentation mocap database hdm05.

[Moeslund et al., 2006] Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2):90–126.

[Muller,] Muller, M. *Information Retrieval for Music and Motion*.

[Müller et al., 2005] Müller, M., Röder, T., and Clausen, M. (2005). Efficient content-based retrieval of motion capture data. *ACM Trans. Graph.*, 24(3):677–685.

[Murthy and Pavlidis, 2005] Murthy, R. and Pavlidis, I. (2005). Non-contact monitoring of breathing function using infrared imaging. Technical Report UH-CS-05-09, University of Houston, TX, 77204, USA.

[Nguyen et al., 2010] Nguyen, D. T., Li, W., and Ogunbona, P. (2010). Human detection using local shape and non-redundant binary patterns. In *11th International Conference on Control, Automation, Robotics and Vision, ICARCV 2010, Singapore, 7-10 December 2010, Proceedings*, pages 1145–1150.

[Nhan and Chau, 2010] Nhan, B. R. and Chau, T. (2010). Classifying affective states using thermal infrared imaging of the human face. *IEEE Transactions on Biomedical Engineering*, 57(4):979–987.

[Ofli et al., 2012] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. (2012). Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. In *CVPR Workshops*, pages 8–13. IEEE.

[Ofli et al., 2013] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. (2013). Berkeley mhad: A comprehensive multimodal human action database. In *WACV*, pages 53–60. IEEE Computer Society.

[Ofli et al., 2014] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. (2014). Sequence of the most informative joints (smij). *J. Vis. Comun. Image Represent.*, 25(1):24–38.

[Ohn Bar and Trivedi, 2013] Ohn Bar, E. and Trivedi, M. (2013). Joint angles similarities and hog2 for action recognition. In *CVPRW*, pages 465–470.

[Ojala et al., 2002] Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987.

[Parameswaran and Chellappa, 2006] Parameswaran, V. and Chellappa, R. (2006). View invariance for human action recognition. *Int. J. Comput. Vision*, 66(1):83–101.

[Pavlidis et al., 2000] Pavlidis, I., Levine, J., and Baukol, P. (2000). Thermal imaging for anxiety detection. In *Proc. IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications*, page 104 –109. IEEE.

[Poppe, 2010] Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vision Comput.*, 28(6):976–990.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *PROCEEDINGS OF THE IEEE*, pages 257–286.

[Raptis et al., 2011] Raptis, M., Kirovski, D., and Hoppe, H. (2011). Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '11, pages 147–156. ACM.

[Raptis et al., 2008] Raptis, M., Wnuk, K., , and Soatto, S. (2008). Flexible dictionaries for action classification. In *The 1st International Workshop on Machine Learning for Vision-based Motion Analysis - MLVMA'08*.

[Romdhani et al., 2006] Romdhani, S., Ho, J., Vetter, T., and Kriegman, D. J. (2006). Face recognition using 3-d models: Pose and illumination. *proce. Of the IEEE*, 294(11):1977–1999.

[Saupe and Vranić, 2001] Saupe, D. and Vranić, D. V. (2001). 3d model retrieval with spherical harmonics and moments. In *DAGM*, pages 392–397. Springer-Verlag.

[Seidenari et al., 2013] Seidenari, L., Varano, V., Berretti, S., Del Bimbo, A., and Pala, P. (2013). Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW '13, pages 479–485. IEEE Computer Society.

[Sempena et al., 2011] Sempena, S., Maulidevi, N. U., and Aryan, P. R. (2011). Human action recognition using dynamic time warping. In Syaichu-Rohman, A., Hamdani, D., Akbar, S., Adiprawita, W., Razali, R., and Sahari, N., editors, *ICEEI*, pages 1–5. IEEE.

[Shimada and Taniguchi, 2008] Shimada, A. and Taniguchi, R.-i. (2008). Gesture recognition using sparse code of hierarchical som. pages 1–4.

[Shotton et al., 2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from a single depth image. In *CVPR*. IEEE.

[Teutsch et al., 2014] Teutsch, M., Muller, T., Huber, M., and Beyerer, J. (2014). Low resolution person detection with a moving thermal infrared camera by hot spot classification. In *CVPR Workshop*.

[Theodorakopoulos et al., 2014] Theodorakopoulos, I., Kastaniotis, D., Economou, G., and Fotopoulos, S. (2014). Pose-based human action recognition via sparse representation in dissimilarity space. *J. Visual Communication and Image Representation*, 25(1):12–23.

[Timo et al., 2004] Timo, A., Abdenour, H., and Matti, P. (2004). Face recognition with local binary patterns. In *ECCV 2004. LNCS*, volume 3021, page 469–481.

[Tran et al., 2008] Tran, D., Sorokin, A., and Forsyth, D. A. (2008). Human activity recognition with metric learning. In *European Conference on Computer Vision*, pages 548–561.

[Turaga et al., 2008] Turaga, P., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Trans. Cir. and Sys. for Video Technol.*, 18(11):1473–1488.

[Tuzel et al., 2006] Tuzel, O., Porikli, F., and Meer, P. (2006). Region covariance: a fast descriptor for detection and classification. In *In ECCV*, pages 589–600.

[Tuzel et al., 2008] Tuzel, O., Porikli, F., and Meer, P. (2008). Pedestrian detection via classification on riemannian manifolds. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 30:1713–1727.

[Vemulapalli et al., 2014] Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. pages 588–595.

[Vieira et al., 2012] Vieira, A. W., Nascimento, E. R., Oliveira, G. L., Liu, Z., and Campos, M. F. M. (2012). Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. 7441:252–259.

[Vranic, 2003] Vranic, D. (2003). An improvement of rotation invariant 3d-shape descriptor based on functions on concentric spheres. pages III: 757–760.

[Wang et al., 2012a] Wang, J., Liu, Z., Chorowski, J., Chen, Z., and Wu, Y. (2012a). Robust 3d action recognition with random occupancy patterns. In *12th European Conference on Computer Vision (ECCV)*, pages 872–885. Springer.

[Wang et al., 2012b] Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012b). Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, pages 1290–1297. IEEE Computer Society.

[Wang and Lee, 2009] Wang, J.-Y. and Lee, H.-M. (2009). Recognition of human actions using motion capture data and support vector machine. *WRI World Congress on Software Engineering, WCSE*, 1:234–238.

[Wang et al., 2010] Wang, W., Zhang, J., and Shen, C. (2010). Improved human detection and classification in thermal images. In *IEEE International Conference on Image Processing (ICIP'10)*, pages 2313–2316. IEEE Press.

[Wanqing et al., 2010] Wanqing, L., Zhengyou, Z., and Zicheng, L. (2010). Action recognition based on a bag of 3d points. In *CVPR Workshop*, pages 9–14.

[Weinkauf et al., 2010] Weinkauf, T., Gingold, Y., and Sorkine, O. (2010). Topology-based smoothing of 2D scalar fields with c1-continuity. volume 29, pages 1221–1230.

[Weinland and Boyer, 2008] Weinland, D. and Boyer, E. (2008). Action recognition using exemplar-based embedding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR'08, June, 2008*, pages 1–7. IEEE.

[Weinland et al., 2011] Weinland, D., Ronfard, R., and Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.*, 115(2):224–241.

[Wong and Cipolla, 2007] Wong, S.-F. and Cipolla, R. (2007). Extracting spatiotemporal interest points using global information. In *ICCV*, pages 1–8. IEEE.

[Xia et al., 2012] Xia, L., Chen, C.-C., and Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. In *CVPR Workshops*, pages 20–27. IEEE.

[Xiaodong et al., 2012] Xiaodong, Y., Chenyang, Z., and Yingli, T. (2012). Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012*, pages 1057–1060.

[Xu et al., 2005] Xu, F., Liu, X., and Fujimura, K. (2005). Pedestrian detection and tracking with night vision. *ITS*, 6(1):63–71.

[Yamato et al., 1992] Yamato, J., Ohya, J., and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pages 379–385.

[Yang and Tian, 2012] Yang, X. and Tian, Y. (2012). Eigenjoints-based action recognition using naïve-bayes-nearest-neighbor. In *CVPR Workshops*, pages 14–19. IEEE.

[Yao et al., 2011] Yao, A., Gall, J., Fanelli, G., and Gool, L. V. (2011). Does human action recognition benefit from pose estimation? In *Proceedings of the British Machine Vision Conference*, pages 67.1–67.11. BMVA Press.

[Yasuno et al., 2004] Yasuno, M., Yasuda, N., and M.Aoki (2004). Pedestrian detection and tracking in far infrared images. In *in Conference on Computer Vision and Pattern Recognition Workshop*, pages 125–131.

[Yilmaz and Shah,] Yilmaz, A. and Shah, M. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *In ICCV*.

[Yoshitomi et al., 1997] Yoshitomi, Y., Miyaura, T., Tomita, S., and Kimura, S. (1997). Face identification using thermal image processing. In *Robot and Human Communication, 1997. RO-MAN '97. Proceedings., 6th IEEE International Workshop on*, page 374–379. IEEE.

[Zatsiorsky, 1998] Zatsiorsky, V. (1998). *Kinematics of Human Motion*. Human Kinetics Publishers, Inc.

[Zhao et al., 2013] Zhao, X., Li, X., Pang, C., and Wang, S. (2013). Human action recognition based on semi-supervised discriminant analysis with global constraint. *Neurocomputing*, 105(Complete):45–50.

[Zhu et al., 2013] Zhu, Y., Chen, W., and Guo, G.-D. (2013). Fusing spatiotemporal features and joints for 3d action recognition. In *IEEE CVPRW*.

[Ziming et al., 2008] Ziming, Z., Yiqun, H., Syin, C., and Liang-Tien, C. (2008). Motion context: A new representation for human action recognition. In

ECCV(4), volume 5305, pages 817–829. Lecture Notes in Computer Science-Springer.